Text-dependent speaker verification: Classifiers, databases and RSR2015

Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li

Text-Dependent Speaker Verification: Classifiers, Databases and RSR2015

Anthony Larcher*, Kong Aik Lee, Bin Ma, Haizhou Li

*Institute for Infocomm Research - A*Star, Singapore*

Abstract

The *RSR2015* database, designed to evaluate text-dependent speaker verification systems under different durations and lexical constraints has been collected and released by the Human Language Technology (HLT) department at Institute for Infocomm Research (I^2R) in Singapore. English speakers were recorded with a balanced diversity of accents commonly found in Singapore. More than 151 hours of speech data were recorded using mobile devices. The pool of speakers consists of 300 participants (143 female and 157 male speakers) between 17 and 42 years old making the *RSR2015* database one of the largest publicly available database targeted for text-dependent speaker verification. We provide evaluation protocol for each of the three parts of the database, together with the results of two speaker verification system: the HiLAM system, based on a three layer acoustic architecture, and an *i*-vector/PLDA system. We thus provide a reference evaluation scheme and a reference performance on *RSR2015* database to the research community. The HiLAM outperforms the state-of-the-art *i*-vector system in most of the scenarios.

Keywords: Speaker recognition, Text-dependent, Database

*Corresponding author

Email addresses: alarcher@i2r.a-star.edu.sg (Anthony Larcher), kalee@i2r.a-star.edu.sg (Kong Aik Lee), mabin@i2r.a-star.edu.sg (Bin Ma), hli@i2r.a-star.edu.sg (Haizhou Li)

[1]alarcher@i2r.a-star.edu.sg

1. Introduction

Speaker verification is the process to accept or reject an identity claim by comparing two speech samples: one that is used as reference of the identity and the other that is collected during the test from the person who makes the claim. Under this generic definition, the claimant is free to provide any utterance for comparison, with no constraints on duration, quality, recording condition and lexical content of the speech sample. The performance of speaker verification suffers from those many possible variabilities of the spoken utterance, amongst which lexical content (Boies et al., 2004; Hébert, 2008) and channel variations (Kinnunen and Li, 2010; Kenny et al., 2007; Wu et al., 2008; Vogt and Sridharan, 2008) are the most detrimental.

It is generally believed that speaker verification achieves better accuracy when the lexical content of the test utterance matches that of the enrollment material, especially in the case of short utterances (Boies et al., 2004; Hébert, 2008). In this regard, two approaches have shown to be effective in tackling the issue of lexical variability. The first approach consist of conducting an *a posteriori* analysis of the speech samples to compensate for the lexical mismatch between enrollment and test utterances (Boakye and Peskin, 2004; Stolcke et al., 2012; Sturim et al., 2002; Vogt et al., 2009) while the second approaches consider the case of cooperative speakers for whom lexical variability can be easily reduced. Despite the higher flexibility of the first approach, it suffers from two drawbacks. On one hand, the lexical analysis increases the computational cost of the verification task. On the other hand, the lexical compensation may be limited by a strong lexical mismatch as it is not possible to guaranty that enrollment and test lexicon overlap. The second approach considers that a cooperative speaker can be asked to pronounce a pre-defined sentence or phrase during both enrollment and test phases. This process is called text-dependent speaker verification as opposed to text-independent speaker verification in which no constraint is put on the input lexicon. In other words, text-dependent speaker verification can be defined as a speaker verification task in which the lexicon used during the test phase is a subset of the lexicon pronounced by the speaker during the enrollment (Hébert, 2008).

Compared to channel variability which usually resulted from uncontrollable environmental factors, lexical variability is relatively more manageable

if we can assume cooperative speakers. With the text-dependent assumption, we achieve a higher accuracy with a shorter duration of both enrollment and test phases by simply forcing the lexical content of the test utterance to match the enrollment material. Therefore, text-dependent speaker verification is well suited for commercial applications which ergonomic constraints require high accuracy and short recording duration.

The absence of lexical constraint on the train/test utterances allows text-independent technique to cover a wide range of applications such as forensic authentication (Campbell et al., 2009; Mandasari et al., 2011), speaker clustering (Silovsky et al., 2011; Brümmer and de Villiers, 2010) and speaker mining (Karam et al., 2011). Moreover, research on text-independent task is strongly supported by the international benchmarking events organized by the National Institute of Standards and Technology (NIST) and the large amount of data that is made available in this context (Martin and Greenberg, 2009). For these reasons, a large scientific community focuses on text-independent speaker verification, despite the commercial potential of text-dependent speaker verification (Lee et al., 2013b; Hébert and Boies, 2005; Wagner et al., 2006; Dialogues Spotlight Technology, 2000; Gu and Thomas, 1998). However, text-dependent speaker verification can be seen as a sub-case of the text-independent task where enrollment and verification utterances have similar duration and lexicon that aim to compensate for the current insufficient performance of more flexible systems. Historically, this statement is supported by a succession of improvements in the field of text-independent speaker verification benefiting the text-dependent sub-case (Schmidt and Gish, 1996; Dong et al., 2008; Aronowitz, 2012; Larcher et al., 2012a).

Recent breakthroughs in terms of accuracy and robustness of text-independent speaker verification systems were achieved at the cost of an intensive use of development data. These improvements have been strongly supported and motivated by the NIST and the Linguistic Data Consortium (LDC[2]) which provide the community with increasingly more challenging data for decades (Martin and Greenberg, 2010). While text-independent speaker verification is using more and more data to train robust systems, research on text-dependency suffers from the lack of data. This leads to practical dif-

[2] `http://www.ldc.upenn.edu/` (Accessed February 28, 2014)

ficulty in adapting existing methods to this specific sub-case. For instance, nine years after its introduction, no paper has been published on the use of Joint Factor Analysis (Kenny and Dumouchel, 2004) for text-dependent speaker verification to our best knowledge. This can partially be explained by the lack of database to support the development of such systems for text-dependent task.

The lack of data affects the text-dependent speaker verification research in more than one way. The limitation of existing databases does not allow a proper study of the effect of lexical variability that would condition the choice of the constraint to put on the speaker. In addition, an overview of existing databases conducted in Section 2 shows imbalanced representation of genders in most of the available corpora when performance of automatic systems are well known to differ across genders (Reynolds et al., 2000; Cumani et al., 2012; Senoussaoui et al., 2011). Finally, the improvement of automatic verification systems calls for a huge number of trials to allow statistically significant performance measures.

In text-dependent speaker verification, the lexical content of the speech data is especially important and there are many ways to constrain the lexicon of the enrollment and test utterances. With different verification protocols, we may need to fix the lexical constraint at different levels such as phone (Matsui and Furui, 1993; Hebert and Heck, 2003), syllable, word (Rosenberg et al., 1991; Kato and Shimizu, 2003) or sentence (BenZeghiba and Bourlard, 2006). Several studies have shown that preserving the lexical sequence within the verification utterances could lead to a 50% relative reduction in terms of error rate (Kato and Shimizu, 2003; Hébert, 2008). Therefore, the choice of a specific protocol is critical from the application point of view as it would strongly affects the accuracy. Nevertheless, very few studies have been conducted to compare the effect of the different lexical constraints (Hébert, 2008), partly due to the lack of databases that could support a fair comparison study.

In this paper, we present the *RSR2015* database that has been released to the public by the Human Language Technology Department[3] at I^{2}R to address some of the limitations of existing corpora (Larcher et al., 2012b).

[3]Institute for Infocomm Research, A*STAR, Singapore, `http://hlt.i2r.a-star.edu.sg/` (Accessed February 28, 2014)

It was recorded as part of the efforts in the deployment of robust speaker recognition for smart-home under the *HOME2015* program (Lee et al., 2011), which leads to its name of *RSR2015* database. The *RSR2015* database is designed to support the research on text-dependent speaker verification and to allow for comparison of verification systems under different lexical and duration constraints. Involving 143 female and 157 male speakers for a total of 151 hours of audio recording, the *RSR2015* database is one of the largest text-dependent speaker verification database publicly available, in terms of speakers and lexical variability. The database is arranged into three parts that address different scenarios. All three parts have been recorded in similar conditions to guarantee a fair comparison between the tasks. The acquisition was realized on six mobile devices including different smart-phones and tablets available commercially. Part I of the *RSR2015* database has been described in (Larcher et al., 2012b).

In the remaining of this paper, we first give an overview of existing databases for text-dependent speaker verification. We summarize 23 databases described in the literature by giving their main characteristics, strengths and weaknesses. The *RSR2015* database is then described in details in Section 3. In the following sections, we propose realistic evaluation protocols and performance measures to allow a fair comparison of systems on the *RSR2015* database. In Section 4, we give a survey of classifiers used for text-dependent speaker verification before describing two state-of-the-art systems that are evaluated on the *RSR2015* database. Section 5 describes the protocols and reports the performance of the two systems on the three parts of *RSR2015* database. Section 6 provides the practical information about how to get this database. We will also discuss some research directions and perspectives regarding text-dependent speaker verification in Section 7.

2. Databases for text-dependent speaker verification

In this section, we present a survey of speech databases available for development and evaluation of text-dependent speaker verification. Although the given list of databases (Table 1) may not be exhaustive, it constitutes the largest inventory in the literature to our best knowledge. Complementary information about resources for speaker recognition can be found in (Campbell and Reynolds, 1999) and a survey of multi-modal biometric databases is given in (Faundez-Zanuy et al., 2006). It is also worth noting that there

Table 1: Overview of existing databases including text-dependent speech material.

Database	Year	Modalities	Ref.	# Speakers Male/Female	Languages	# Sessions	Environment	Intersession Interval	Age Info
YOHO	1995	Sp	Campbell and Higgins (1994); Campbell (1995)	138 106/32	EN	14	Quiet	3 days	no
BT-DAVID	1996	Sp,Vi	Mason et al. (1996)	31 15/16	EN	5	Quiet	days/months	yes
M2VTS	1997	Sp,2Fa	Pigeon and Vandendorpe (1997)	37 30/7	EN	5	Quiet	1 week	no
PolyVAR	1997	Sp	Chollet et al. (1996)	143 85/58	EN	1-229	Quiet	days/months	yes
OGI Speaker Recognition	1998	Sp	Cole et al. (1998)	91 43/48	EN	12	Quiet/Noisy	months/years	yes
XM2VTS	1999	Sp,Vi	Messer et al. (1999)	295 158/137	EN	4	Quiet	weeks/months	no
Ahumada	2000	Sp	Ortega-Garcia et al. (2000)	104 104/0	SP	6	Quiet	> 11days	yes
PolyCOST	2000	Sp	Hennebert et al. (2000)	134 74/60	EN,EU	5-14	Quiet	3 days	yes
Verivox	2000	Sp	Karlsson (1999); Karlsson et al. (2000)	50 50/0	SW	2	Quiet	same day	no
SmartKom	2002	Sp,Ir,Vi	Steininger et al. (2002)	45 20/25	GE	2	Quiet	same day	no
BANCA	2003	Sp,Vi	Bailly-Bailliere et al. (2003)	208 104/104	EN,FR,IT,SP	12	Quiet/Noisy	-	no
BIOMET	2003	Sp,2Fa,3Fa,Fp,Hg,Sg	Garcia-Salicetti et al. (2003)	91 45/46	FR	3	Quiet	months	yes
STC	2003	Sp	ELDA - Evaluations and Language resources Distribution Agency (2003)	89 54/35	RU	1-15	Quiet	months	no
MyIdea	2005	Sp,Fp,Hg,Pp,Sg,Vi	Dumas et al. (2005)	30 30/0	EN.FR	3	Quiet/Noisy	days/months	no
Valid	2005	Sp,Vi	Fox et al. (2005)	106 76/30	EN	5	Quiet	weeks	yes
CCC-VPR2C2005 10000	2006	Sp	Zheng (2005)	10,000 -/-	PU	2	Quiet	-	no
MIT-MDSVC	2006	Sp	Woo et al. (2006)	88 49/39	EN	2	Quiet/Noisy	days	no
M3	2006	Sp,2Fa,Fp	Meng et al. (2006)	39 29/10	CA,EN,PU	3	Quiet/Noisy	months	yes
BIOSEC	2007	Sp,2Fa,Fp,Is	Fierrez et al. (2007); Toledano et al. (2008)	250 -/-	EN,SP	4	Quiet	months	yes
BioSecurID	2007	Sp,2Fa,Fp,Hg,Is,Ks,Pp,Sg	Fierrez et al. (2010)	400 -/-	SP	4	Quiet	months	yes
MBioID	2007	Sp,2Fa,3Fa,Fp,Is,Sg	Dessimoz et al. (2008)	120 -/-	EN,FR	2	Quiet	same day	yes
BioSecure	2010	Sp,Fp,Hg,Is,Sg,Vi	Ortega-Garcia et al. (2010)	400 -/-	EU	2	Quiet	months	yes
UNMC-VIER	2011	Sp,Vi	Wong et al. (2011)	123 74/49	EN	2	Quiet	same day	no
RSR2015	2012	Sp	Larcher et al. (2012b)	300 157/143	EN	9	Quiet	same day	yes

The nomenclature for biometric modalities is as follows: 2Fa stands for Face 2D, 3Fa stands for Face 3D, Fp stands for Fingerprint, Hg stands for Hand geometry, Ir stands for Infra-red video, Is stands for Iris, Ks stands for Keystrokes, Pp stands for Palm-print, Sg stands for Handwritten signature, Sp stands for speech and Vi stands for Video.
Languages abbreviations are used as follows: CA stands for Cantonese, EN stands for English, EU stands for various European languages, FR stands for French, GE stands for German, IT stands for Italian, PU stands for Putonghua, RU stands for Russian, SP stands for Spanish and SW stands for Swedish.
The number specified in column 5 corresponds to the number of speakers who completed all recordings for enrollment and test sessions given the standard protocol released with the database. Additional recordings for impostor speakers are distributed with some of the databases (e.g. BT-DAVID, M3, Biosecure).

have been some reported results i the literature on databases that are not publicly available (Li et al., 2002; Toledo-Ronen et al., 2011).

Our intention is to provide some context about the motivations of the *RSR2015* database but not to give an exhaustive description of existing databases. Indeed, databases for text-dependent speaker verification have been designed for various purposes and the diversity of protocols makes it difficult for a fair comparison of the corpora. In the remaining of this section, we discuss some of the main characteristics of the existing databases related to the major challenges of text-dependent speaker verification.

2.1. Demography

Population demographics are critically important when evaluating the performance of speaker verification systems (Doddington, 2012). In speaker verification, where development and evaluation of automatic systems are driven by existing corpora, the population recorded for a database has to be carefully selected. For specific applications, the population is selected to be as representative as possible of the target population whereas databases designed for generic research purpose tend to cover the largest possible population. In the remaining of this paper we consider the demography of the population in terms of gender and age which are often considered as two of the main criteria affecting speaker verification engines. For this reason, the set of recorded speakers should be representative of the gender and age distribution of the target population. Additionally, the population needs to be large enough as improvement in performance of automatic speaker verification systems requires enormous number of trials to ensure the results are statistically significant (Doddington, 1998).

In practice, the size of the population, together with its representativeness are limited by technical and pecuniary concerns. Interestingly, a large part of the databases that include text-dependent speech material are multi-modal databases, i.e., out of the 24 databases listed in Table 1, 14 are multi-modal. The advantage of collecting multi-modal databases is twofold. First it allows research on comparison and combination of modalities for person authentication (Marcel et al., 2010). Second, it pools the cost and complexity load that goes along the collection of biometric samples. The huge effort that the scientific community has put in collecting data to sustain the research on biometrics in the past twenty years can be acknowledged from Table 1. Nevertheless, the number of speakers enrolled in those database is still limited as only 7 of the 24 entries in the table count more than 200 subjects.

Another limitation is the imbalanced gender representation that can be observed from Figure 1. Out of the 19 databases for which the gender information is available, 8 can be considered as gender balanced with at least 45% of speakers for each gender while 7 of the databases include less than 30% of female speakers. This disequilibrium is especially damaging as the performance of speaker verification systems is known to differ for male and female speakers (Doddington, 2012). Furthermore, information about the age is not always available (at least in the documentations publicly available and listed in Table 1). The discrimination between speakers has been shown to be more difficult when the age difference is small (Doddington, 2012).

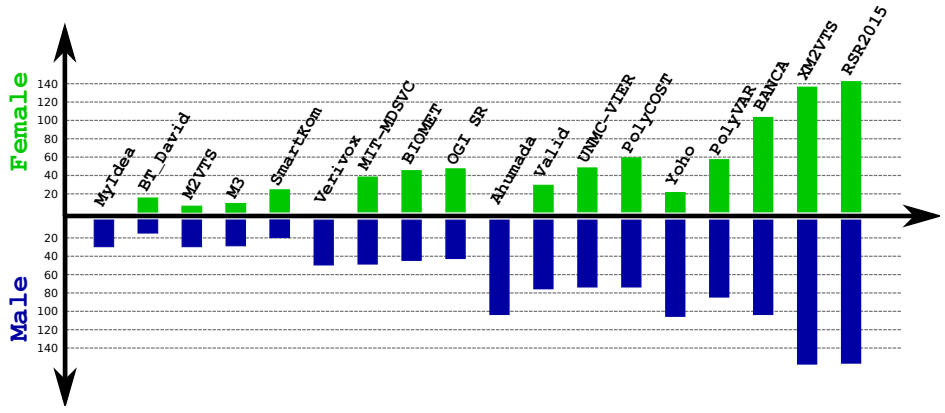


Figure 1: Comparison of the number of speakers per gender in publicly available databases for text-dependent speaker recognition. Only databases for which the gender information is available have been sorted by ascendant total number of speakers.

2.2. Lexical variability

Performances of speaker verification systems are known to be strongly dependent on the condition of the speech material provided as input. For instance, many studies have been carried out to estimate the impact of speech duration (Vogt et al., 2008; Fauve, 2009; Kanagasundaram et al., 2011). Other works have shown that discriminancy depends on the speech contents that were used for enrollment and test, leading to the conclusion that, for a fixed duration, different parts of an utterance might not be equally useful for speaker verification (Amino and Arai, 2009; Kahn et al., 2010; Nosratighods et al., 2010; Kahn et al., 2011). In text-dependent speaker verification where both enrollment and test utterances are fixed, lexical content is especially important as it can affect the accuracy of the system (Kato and Shimizu, 2003; Hébert, 2008). Thus, influence of the selected lexical content should be studied when deploying a text-dependent speaker verification system.

2.2.1. Main stream protocols for speaker verification

In the past twenty years, large databases and their associated protocols provided by the NIST have become a de facto standard for the evaluation of text-independent speaker verification technologies (Martin and Greenberg, 2009). No such standard exists for the case of text-dependent speaker verification, making the comparison across systems difficult and multiplying

the number of protocols reported in the literature. Nevertheless, two main streams are reflected in the existing databases. In Yoho, M2VTS, Verivox and Biosec, the lexical contents of the training and test utterances are strongly constrained by using only digits, while databases such as SmartKom, STC, CCC-VPR2C2005-10000 or MIT-MDSVC allow a wider lexical coverage by using fixed phrases. Those two types of protocols are covered by databases such as Polyvar, OGI speaker verification, XM2VTS, Ahumada, PolyCost, BANCA, BioMet, MyIdea, Valid, M3, BiosecureID, MBioID, BioSecure or UNMC-VIER which offer different sets of digits strings together with fixed phrases. Most of the time, the lexical variability is limited to a few fixed sentences and fixed digit strings. For instance, out of the 24 listed in Table 1, 10 databases contain less than 10 different sentences.

In order to increase the lexical coverage, some databases like PolyCost, Banca, MyIdea or MIT-SDSVC include lexical content that varies across speakers. Trials in which the impostor pronounces the text used by the target speaker to enroll are produced by asking each subject to pronounce the content of some other subjects. Under such protocol, the possibilities of inter-speaker impostor trials are greatly limited as the impostors and target don't all speak the same speech content.

2.2.2. Languages for text-dependent speaker verification

As the lexical content is constrained by the language of the application (Li et al., 2013) an important effort has been observed in the recent years to provide the community with resources in languages such as French (Bailly-Bailliere et al., 2003; Garcia-Salicetti et al., 2003; Dumas et al., 2005; Dessimoz et al., 2008), German (Steininger et al., 2002), Italian (Bailly-Bailliere et al., 2003), Mandarin Chinese (Zheng, 2005; Meng et al., 2006), Russian (ELDA - Evaluations and Language resources Distribution Agency, 2003), Spanish (Ortega-Garcia et al., 2000; Bailly-Bailliere et al., 2003; Dessimoz et al., 2008; Ortega-Garcia et al., 2010) or Swedish (Karlsson, 1999). A few databases, mostly due to collaborative efforts within the European Union (Hennebert et al., 2000; Bailly-Bailliere et al., 2003; Ortega-Garcia et al., 2010), also include multi-lingual contents. However, 10 databases out of the 24 listed in Table 1 contain only English speech when another 6 include English contents in addition to another language. The omnipresence of English in the existing protocol is mainly due to the fact that English speakers are easily available in addition to the local ones (Dumas et al., 2005; Meng et al., 2006; Fierrez et al., 2007; Toledano et al., 2008; Dessimoz et al., 2008) or that

English is used as an international standard for historical reasons.

2.3. Session variability

The mismatch between enrollment and test utterances can be greatly reduced by matching the lexical content of both utterances, making the speaker verification task easier when dealing with short duration (Hébert, 2008). Nevertheless, other factors that we refer to as session variability still affect the performance of speaker verification systems such as channel mismatch, ambient noise or intra-speaker variability. In the remaining of this article, the term session is used to refer to recordings that differ by one or more element such as environment, recording device or time.

Due to the complexity and the cost of data acquisition, especially for the case of multi-modal corpora, most databases were recorded using the same microphone and under controlled environment,which strongly limits the channel and noise variability across sessions (e.g. Yoho, BT-David, M2VTS, XM2VTS, Verivox, SmartKom, Biomet, STC, Biosec). Other databases focus especially on adverse condition by providing recordings of speakers in various environments such as outdoor, in the street, in a public area like building lobby or cafeteria (e.g. BANCA, MyIdea, MIT-MDSVC or M3). Those databases are labeled as *Noisy* in the eighth column of Table 1 in contrast to to other databases that do not explicitly address environment mismatch. Finally, some databases include explicit channel mismatch with speakers recorded on different devices but do not impose any background noise or environment factors during the recording (e.g. PolyCOST or PolyVAR).

The number of sessions in text-dependent databases is often limited due to the cost of recording which is proportional to the number of times and the duration on which a speaker has to be mobilized. Amongst the 24 databases listed in Table 1, 16 include less than 5 sessions per speaker (Table 1). A number of databases have been recorded with a special attention to the time interval between two sessions (e.g., OGI Speaker Recognition, Ahumada, Biomet, Valid, etc.) to maximize the within speaker variability as influence of aging is well known in biometrics. However, (Lei and Hansen, 2009; Lawson et al., 2009; Kelly and Harte, 2011; Kelly et al., 2012) show that, for the case of text-independent speaker verification, aging effect only becomes significant after a period of several years that is only covered by the OGI Speaker Recognition database (or by the Greybeard database for the

case of text-independent speaker recognition[4]). On the contrary, other works show significant degradation appearing after a period of months for the case of text-dependent speaker verification (Furui, 1981b; Mistretta and Farrell, 1998). Those studies suggest that aging phenomenon is not well understood yet and might differently affect text-dependent and text-independent speaker verification.

3. *RSR2015* database

The *RSR2015* database was recorded in order to provide the community with a sufficiently large dataset from a gender-balanced set of speakers. It consists of recordings from 300 speakers in 9 sessions recorded with multiple hand-phones and tablets. The 196,844 files resulting from this recording contain 151 hours and 30 minutes of audio signal.

A special attention has been paid to the lexical content in order to allow for fair comparison of speaker verification systems under different lexical constraints. Therefore, recordings of the 300 speakers are divided into three parts, each dedicated to a specific task involving different lexical and duration constraints. Part I of the *RSR2015* database is dedicated to speaker verification using fixed short pass-phrases. Part II is dedicated to speaker-loaded command control (Lee et al., 2011). Part III is devoted to speaker verification using randomly prompted digit strings. To allow a fair comparison between use-cases, the three parts have been recorded in similar conditions with the same speakers and channels.

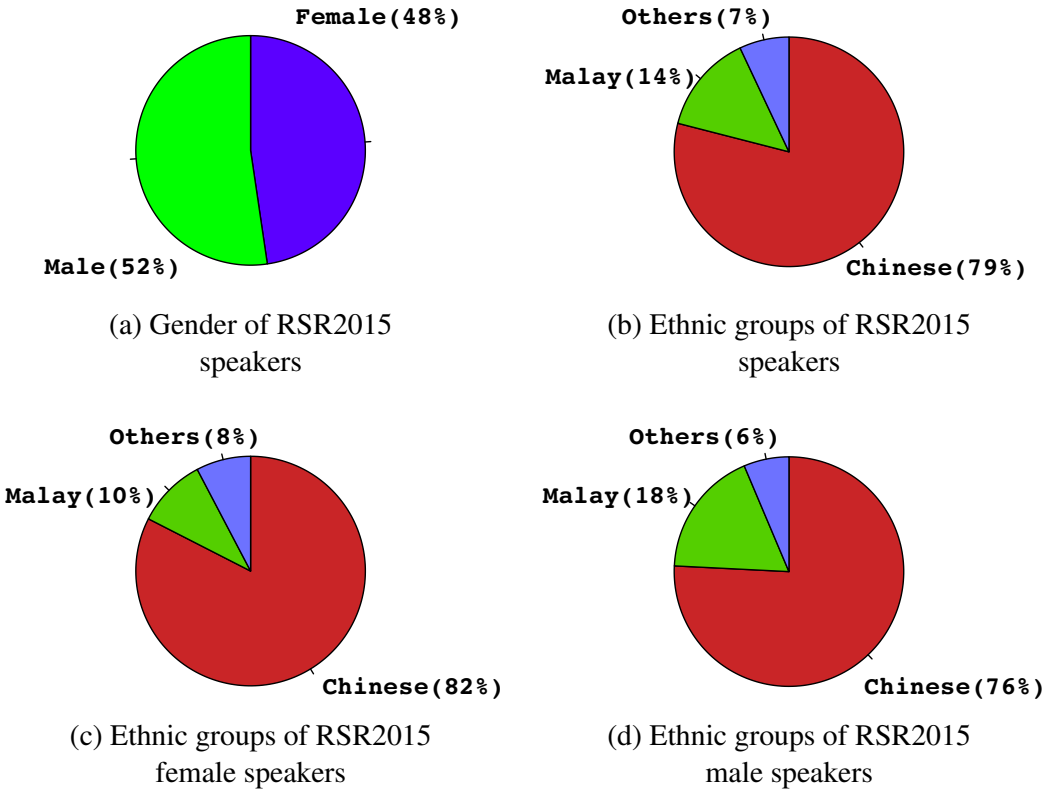
3.1. Demography

With 300 speakers, the *RSR2015* database is one of the largest database publicly available for text-dependent speaker verification. To our best knowledge, the only databases including more speakers for text-dependent speaker verification are BioSecure, BioSecurID and the CCC-VPR2C2005-10000 which lexical content is narrower than *RSR2015* database (Section 3.3).

In order to be representative of the Singaporean population, the 300 speakers recorded for the *RSR2015* database have been selected according to their ethnic group and gender. The result is a gender-balanced database in which 143 speakers out of the 300 are female (Figure 2(a)). Additionally, 237

[4]LDC Catalog No LDC2013S05

speakers are of Chinese origin, 42 are of Malay origin while the remaining 21 are from other various ethnic groups (Figure 2(b), 2(c) and 2(d)).



(a) Gender of RSR2015 speakers

(b) Ethnic groups of RSR2015 speakers

(c) Ethnic groups of RSR2015 female speakers

(d) Ethnic groups of RSR2015 male speakers

Figure 2: Gender and ethnic statistics from the 300 speakers of the *RSR2015* database. A special attention has been ported to balance genders (female: 143, male: 157) and ethnic origins to reflect Singapore population (Chinese: 237, Malay: 42 and others: 21).

The *RSR2015* database includes speakers from 17 to 42 years old (Figure 3). Given the limited sample size of the *RSR2015* database, widening the age bracket would create a sparse distribution of speakers across ages that may artificially facilitate the task of speaker verification. Indeed, it was shown in (Doddington, 2012) that the difficulty of the speaker verification task increases when the age difference between speakers is limited. Therefore a population of speakers in a limited age bracket may increase the challenge of speaker verification.

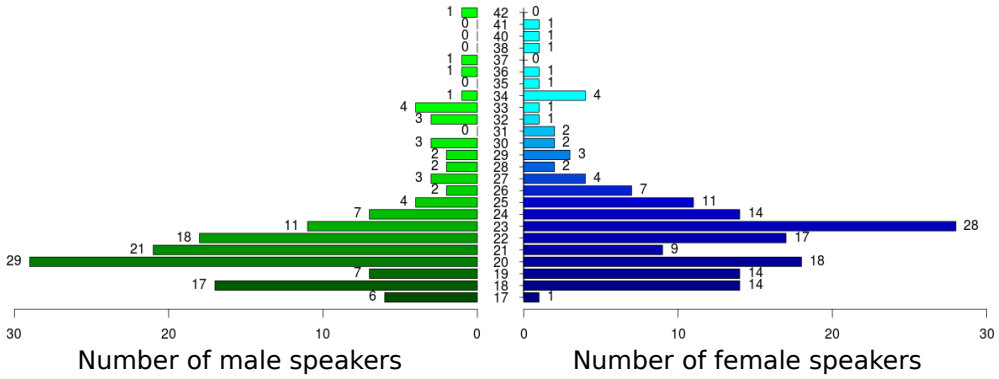


Figure 3: Age pyramid of both male and female speakers of the *RSR2015* database.

3.2. Acquisition Protocol

The nine sessions of the *RSR2015* database were recorded indoor under a typical office environment. Each subject completed the recording process on a single day so the *RSR2015* database does not include aging variability. However, it has been shown in (Lawson et al., 2009) that aging variability within 3 years is negligible compared to session variability.

Six mobile devices[5] (five smart-phones and one tablet) available in the market were used for recording. Three portable devices (labeled A, B and C) were assigned to each subject. The nine sessions of each subject were then recorded using the three devices in the following sequence: $\{A,B,C,A,B,C,A,B,C\}$ and the meta-data information is provided together with the data. A dialogue manager was implemented as an Android© application to manage the recording. This application uses the touch-screen capability of the devices to prompt the text content. A push-to-talk feature was used to allow the user to start the recording and stop it after reading the prompt. The subject was free to hold the portable device in a way (s)he was comfortable and acoustic quality can thus vary significantly within and across sessions.

The audio signal was recorded through the internal microphone of each of the six portable devices in raw PCM format, at 16 kHz sampling frequency with a resolution of 16 bits per sample. A SPHERE[6] header was added afterwards to each file including information about the device, the language,

[5]Samsung Nexus©, Samsung GalaxiS© ×2, HTC Desire©, Samsung Tab©, HTC Legend©

[6]`http://www.itl.nist.gov/iad/mig/tools/` (Accessed February 28, 2014)

the number of samples, the sample rate and the data format.

3.3. Three different text-dependent tasks

Different scenarios could be used to constrain the lexical content of the enrollment and test utterances. Despite the different lexical and duration constraints required by various scenarios, existing databases rarely include data that allow comparison of systems across scenarios. Only 8 databases out of the 24 listed in Table 1 include material that can be used to study the co-articulation effect which strongly affects the performance of verification systems (Kato and Shimizu, 2003). In order to allow comparison of systems across different scenarios, the three parts of the *RSR2015* database have been designed with different lexical constraints[7] while keeping identical recording conditions. For all three parts described below, all 300 speakers pronounce the same lexical content within a given session. In each session, a given speaker pronounces each sentence exactly once.

Part I of the *RSR2015* database focuses on a text-dependent speaker verification task where speakers pronounce fixed pass-phrases to authenticate. In each of the nine sessions, a speaker pronounces 30 fixed sentences selected from the TIMIT database (Garofolo et al., 1993) to cover all English phonemes. The average recording duration across speakers, sessions and sentences is 3.20 seconds and the average duration per sentence varies from 2.73 to 3.65 seconds. Note that these sentences have been selected to evaluate the impact of different lexical content with a similar duration. After applying the energy-based speech activity detection (SAD) as described in Section 5.1, the average nominal speech duration across sentences is 1.25 seconds[8] (varying from 1.01 to 1.59 seconds across sentences). The entire Part I of the *RSR2015* database consists of 72 hours of audio recording (28 hours and 15 minutes of nominal speech after SAD).

Part II of the *RSR2015* database focuses on a speaker-loaded command control task where speakers pronounce fixed commands to control home

[7] described in Appendix Appendix A

[8] Drastic duration reduction after applying SAD is partly due to silence removal before and after the utterance as the recording was controlled by the speakers through a push-to-talk process.

appliances and be authenticated at the same time. In each of the nine sessions, a speaker pronounces 30 short commands defined to control home appliances of the *StarHome*, a fully functional 180 square meters smart home prototype located at the Fusionopolis, Singapore (Lee et al., 2011). Average recording duration across speakers, sessions and commands is 1.99 seconds and average per command duration vary from 1.66 to 2.46 seconds. After applying the energy-based SAD, the average nominal speech duration across commands is 0.63 seconds (varying from 0.44 to 0.99 seconds across sentences). The entire Part II of the *RSR2015* database consists of 44 hours and 53 minutes of audio recording (14 hours and 12 minutes of nominal speech after SAD).

Part III of the *RSR2015* database focuses on a text-dependent speaker verification task where speakers are prompted with random sequences of digits. In each of the nine sessions, a speaker pronounces 3 sequences of ten digits and 10 sequences of five digits. The digit sequences are different across sessions but identical for all speakers in order to generate trials where impostor pronounce the correct sequence. The speech material used for enrollment and test is constrained to ten English digits (zero - one - two - three - four - five - six - seven - eight - nine) but the left-right context of each digit is different between enrollment and test in order to evaluate the effect of co-articulation.

For the ten-digit sequences, the average recording duration across speakers, sessions and sequences is 5.19 seconds. After applying the energy-based SAD, the average nominal speech duration across sequences is 2.07 seconds. For the five-digit sequences, the average recording duration across speakers, sessions and sequences is 3.06 seconds. After applying the energy-based SAD, the average nominal speech duration across sequences is 1.09 seconds.

The entire Part III of the *RSR2015* database consists of 34 hours and 36 minutes of audio recording (12 hours and 51 minutes of nominal speech after SAD).

4. Classifiers for text-dependent speaker verification

Meaningful comparison of accuracy in text-dependent speaker verification tends to be very difficult due to the lack of standard evaluation protocol and

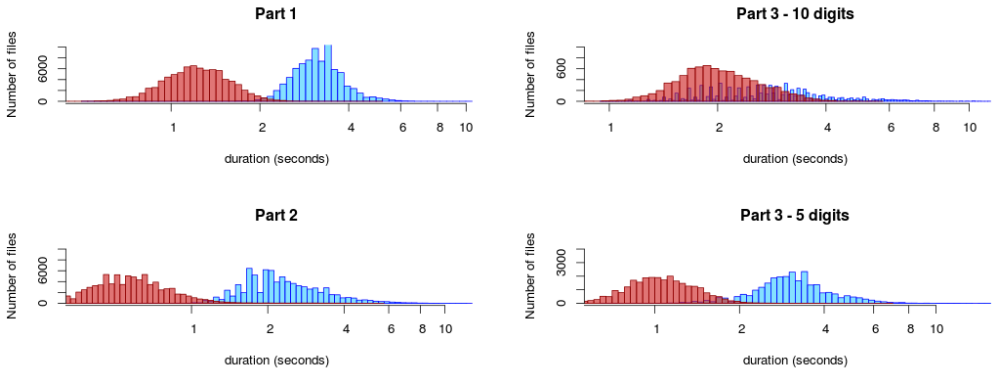


Figure 4: Distribution of the audio recording (in light blue) and nominal speech duration (in dark red) in seconds for the three parts of the *RSR2015* database. For Part III, durations are given for 10-digits sequences and 5-digits sequences separately.

database which motivates the collection of the *RSR2015* database. Moreover, system architectures can be strongly influenced by specific use-cases and their inherent lexical constraints. In this section, we first give an overview of existing classifiers used for text-dependent speaker verification. We describe then two state-of-the-art speaker verification systems: a text-dependent system based on GMM and HMM modeling and an i-vector system, which performance on the *RSR2015* database is given as baseline in Section 5. The choice of these classifiers is motivated by their representativeness of current speaker verification engines. Indeed, the text-dependent engine has been recently deployed in a large-scale commercial application while the i-vector system is adapted from the main-stream state-of-the-art engines for text-independent speaker verification (Dehak et al., 2011a). To establish a fair comparison, both systems use the same front-end and their performance is given by using two performance metrics described thereafter.

4.1. Survey of existing classifiers

A specificity of text-dependent speaker verification systems is that they have to model the speaker characteristics together with the lexical content of the verification utterances. In the last thirty years, two major trends have been dominating the field of text-dependent speaker verification.

A first category of classifiers, based on dynamic programming has been proposed when the quantity of speech available is limited (Furui, 1981a; Dutta, 2008). Working at the frame level, they offer a precise modeling of the

temporal structure of the speech utterances but lack the generalization power offered by generative approaches. In particular, as the dynamic programming technique typically provides Euclidean distance rather than likelihood probability between speech samples, thus requiring additional decision mechanism for speaker verification which is typically formulated as a hypothesis test in probabilistic domain. Several attempts have been made to compensate for the intra-speaker variability by introducing a distance normalization (Luan et al., 2006) or a multi-template approach (Ramasubramanian et al., 2006). Additional information such as suprasegmental and source features can also be used to reinforce the system (Yegnanarayana et al., 2005; Avinash et al., 2010).

A second category of classifiers, by far the most common, is based on Hidden Markov Models (HMMs). HMMs are inherently more robust to the variability of the speech signal and can take advantage of a larger quantity or enrollment data. Additionally, they benefit from the progress achieved in the fields of text-independent speaker verification (Kinnunen and Li, 2010) and speech recognition (Young, 2008). In practice, text-dependent speaker verification faces different use cases, each of which has a unique modeling and run-time requirement. With HMM, granularity of models can be tailor-made to represent the temporal structure of the speech utterances. Systems based on phone models offer the finest granularity and thus can be used for any lexical content (Matsui and Furui, 1993; Che et al., 1996; Charlet and Jouvet, 1997; Nakagawa et al., 2004) while HMMs modeling words (Rosenberg et al., 1991; Yoma and Pegoraro, 2002; Kato and Shimizu, 2003) or entire utterances (Rosenberg et al., 2000; Forsyth, 1995; Subramanya et al., 2007; Charlet et al., 2000; Larcher et al., 2013b), which granularity is less, are restrained to limited lexicon. Research is also carried out to improve the robustness of such models to channel and speaker variability. In (Chatzis and Varvarigou, 2007), the Gaussian distributions of the HMMs states are replaced by Student-t distributions, more robust to noise. In (Aronowitz, 2012), the authors adapt the concept of support vector machines together with the nuisance attribute projection (NAP) (You et al., 2010) to be used with HMMs. Despite the good performance of this approach, it is limited to the case where all users of the system share the same pass-phrase, due to the amount of data required to train the NAP matrix.

Other works in the literature propose to model the temporal structure of the speech utterance by using artificial neural network (Chen et al., 1996; Finan et al., 1996; Woo et al., 2000) or make use of spectrogram-based rep-

resentation (Das and Tapaswi, 2010; Dutta, 2007; Kekre et al., 2010). The different modeling approaches can eventually be combined in order to compensate for individual weaknesses (Farrell, 1995; Farrell et al., 1998; Bonastre et al., 2003).

Finally, progress of text-independent speaker verification have inspired a number of systems. Architectures based on the classical GMM/UBM (Boies et al., 2004; Aronowitz, 2012; Chen et al., 2012; Hebert and Heck, 2003) or on the more recent i-vector representation (Aronowitz, 2012; Larcher et al., 2012a, 2013c; Stafylakis et al., 2013) have been shown to take advantages of the lexical information required by text-dependent speaker verification. These systems have achieved a limited success as they don't explicitly take advantages of the temporal structure of the speech utterances. It is however possible to combine these approaches with a speech recognition engine to effectively verify both speaker and speech content (Heck and Genoud, 2001) in exchange for an extra computational cost.

From a Bayesian perspective, systems based on generative approaches are superior to those relying on dynamic programing in the sense that they can produce likelihood ratio scores, easier to calibrate and interpret when taking a decision (van Leeuwen and Brümmer, 2013). In the following, we present two state-of-the-art speaker verification systems: the text-dependent HiLAM speaker verification engine based on GMM and HMM modeling and an i-vector system, which both produce natural likelihood ratios. The HiLAM has been extensively tested (Lee et al., 2013b) for its robustness and practicality in commercial deployments, and the i-vector system is derived from the state-of-the-art text-independent speaker verification engines (Larcher et al., 2013b).

4.2. Text-dependent system: HiLAM

4.2.1. Utterance modeling

The Hierarchical multi-Layer Acoustic Model (HiLAM) is a text-dependent speaker verification engine that has been described in (Lee et al., 2011; Larcher et al., 2012b). It is an extension of the classical GMM/UBM approach as depicted in Figure 5. All the emission probabilities in this architecture are mixtures Gaussian distributions (GMM) sharing the same variance and weight parameters. The first two layers are similar to the standard GMM/UBM in which the UBM at the upper layer models the general speech acoustic space. The middle layer is the text-independent speaker model ob-

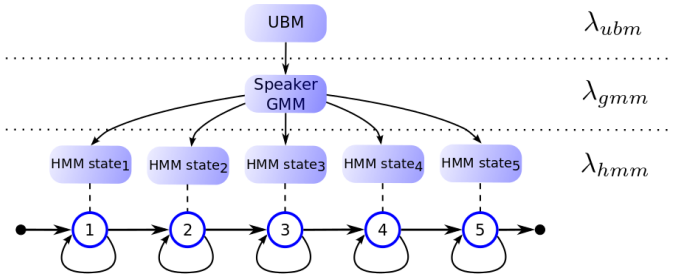


Figure 5: The Hierarchical multi-Layer Acoustic Model (HiLAM).

tained by a classical Maximum *a Posteriori* (MAP) adaptation of the UBM. The bottom layer hinges on the abilities of a left-to-right Hidden Markov Model (HMM) to harness the specific temporal structure of pass-phrases. The emission probability density function of each HMM state is derived from the middle-layer speaker-dependent GMM. Each of those GMMs is adapted from the text-independent speaker model following the MAP criterion. Only the mean parameters are adapted, which is different from that proposed originally in (Larcher et al., 2008) where the weights parameters are adapted. This essentially replaces the semi-continuous HMM (SCHMM) (Young, 1992) used in the original work with a continuous density HMM (CDHMM) providing higher accuracy at the expense of higher computation.

The training of the HiLAM is similar to the original one described in (Larcher et al., 2008). A gender-independent UBM is firstly trained to model the acoustic space. The text-independent speaker model is then adapted from the UBM with all data pronounced by the target speaker. Finally an iterative training is performed to train the third layer's HMM. In order to initialize the HMM for each pass-phrase, the utterance is cut into S segments $\{seg_i\}_{i\in[1,S]}$ of the same length. Each state of the HMM is adapted from the middle layer GMM using the corresponding seg_i. A new segmentation is then performed using the adapted HMM. Viterbi algorithm is used for this purpose. This iterative process is performed until convergence of the Viterbi path. The number of states S is chosen empirically. Transitions of the left-to-right HMM are set equiprobable.

During testing, given a speech sequence X, a text-dependent score, $S_{TD}(X)$, is computed as:

$$S_{TD}(X) = \log \frac{\mathcal{L}_{HMM}(X)}{\mathcal{L}_{UBM}(X)} \tag{1}$$

where $S_{TD}(X)$ is the log-ratio between the likelihood of the given sequence over the speaker's text-dependent HMM aligned by Viterbi decoding, $\mathcal{L}_{HMM}(X)$, and the likelihood of X on the UBM, $\mathcal{L}_{UBM}(X)$. The number of states for each semi-continuous HMM is empirically set to 5 when modeling sentences from the Part I and it is set to 3 when modeling the shorter commands from Part II.

4.2.2. Digit modeling

A modified version of the HiLAM has been developed to deal with randomly prompted digits. The two upper layers of the architecture are similar to the original model. During the enrollment, each speaker pronounces several occurrences of the ten English digits. Recordings from the target speaker are automatically segmented to train a set of ten GMMs (one per digit) by adapting the speaker-dependent GMM from the middle layer using a Maximum a Posteriori (MAP) criterion. Note that the segmentation of the enrollment utterances is done using a state-of-the-art speech recognition system and thus no iterative adaptation is performed to train the HMM components. During testing, given a randomly prompted sequence of N digits, a left-to-right HMM is composed with the corresponding N digit models. The verification score is then computed according to Equation 1 where the likelihood of the test segment over the HMM is obtained using a Viterbi alignment.

4.3. Standard i-vector system

The paradigm of i-vectors (Dehak et al., 2011a) is based on the assumption that speech segments of variable duration can be represented as fixed dimension vectors, the i-vectors, in a low-dimensional space referred to as total variability space. Taking advantage of the low dimensionality of the total variability space, many classifications techniques have been applied to perform different tasks such as speaker and language recognition (Dehak et al., 2011b; Bousquet et al., 2011; Kanagasundaram et al., 2011; Mandasari et al., 2011; Xu et al., 2011) or speaker diarization (Prazak and Silovsky, 2011). As i-vectors retain different types of variability available in the speech segments, such as speaker and lexical content, recent works have shown that i-vectors can be used for the task of text-dependent speaker recognition (Larcher et al., 2012a; Aronowitz, 2012; Larcher et al., 2013c).

4.3.1. I-vector extraction and normalization

The projection of a speech segment onto the total variability space can be considered as a probabilistic compression process that reduces the dimensionality of a channel- and speaker-dependent super-vector of concatenated Gaussian Mixture Model (GMM) means, according to a linear-Gaussian model. The super-vector, $\mathbf{m}$, is projected onto the total variability space according to the generative equation:

$$\mathbf{m} = \boldsymbol{\mathcal{M}} + \mathbf{T}\boldsymbol{\phi} \tag{2}$$

where $\boldsymbol{\mathcal{M}}$ is a speaker and channel independent super-vector, $\mathbf{T}$ is a factor-loading low-rank matrix and $\boldsymbol{\phi}$ is a random vector that is assumed to follow a standard normal distribution. An i-vector $\mathbf{x}$ is the maximum a posteriori point estimate of $\boldsymbol{\phi}$ given a speech utterance. More details about the i-vector extraction process can be found in (Dehak et al., 2011a; Martinez et al., 2011).

Most of the classification techniques using i-vector assume that they follow a Gaussian distribution which is not the case in practice. Several normalization algorithms have been proposed to modify the i-vector distribution according to the Gaussian assumption (Dehak et al., 2010; Bousquet et al., 2011; Garcia-Romero and Espy-Wilson, 2011). Spherical Nuisance Normalization, *SphNorm*, has been shown to produce good performance when associated with Probabilistic Linear Discriminant Analysis (PLDA)(Bousquet et al., 2012).

SphNorm is an iterative process which parameters are estimated in a large development set of i-vectors. For each iteration n the mean $\boldsymbol{\mu}_n$ and within-class covariance $\mathbf{W}_n$ of the development set are computed. All i-vectors $\mathbf{x}$ from the development set are then normalized according to the following algorithm:

Spherical Nuisance Normalization algorithm for *i*-vector normalization

Given a test vector $\mathbf{x}$,

for $n = 1$ to *nb_iterations*: $\quad \mathbf{x} \leftarrow \dfrac{\mathbf{W}_n^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}_n)}{\left\|\mathbf{W}_n^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}_n)\right\|}$

I-vectors from the test set are then normalized following the same transformation.

4.3.2. Speaker Modeling with Probabilistic Linear Discriminant Analysis

Introduced in (Prince and Elder, 2007), PLDA is a generative model which assumes that the observation $\mathbf{x}_{i,j}$ of a speaker i in a session j is a sum of four components

$$\mathbf{x}_{i,j} = \boldsymbol{\mu} + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{i,j} + \boldsymbol{\epsilon}_{i,j} \tag{3}$$

where $\boldsymbol{\mu}$ is the mean of the i-vector distribution, $\mathbf{F}$ and $\mathbf{G}$ are low rank matrices which column vectors form bases of two sub-spaces that are supposed to contains the speaker and session variability respectively. Thus, $\mathbf{h}_i$ and $\mathbf{w}_{i,j}$ are latent variables related to $\mathbf{x}_{i,j}$ in these sub-spaces. $\boldsymbol{\epsilon}$ is a normally distributed additive noise of covariance matrix $\boldsymbol{\Sigma}$ and conditional and prior densities are given by:

$$\begin{aligned} P(\mathbf{x}_{i,j}|\mathbf{h}_i, \mathbf{w}_{i,j}) &= \mathcal{N}_{\mathbf{x}_{i,j}}(\boldsymbol{\mu} + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{i,j}, \boldsymbol{\Sigma}) \\ P(\mathbf{h}_i) &= \mathcal{N}_{\mathbf{h}_i}(\mathbf{0}, \mathbf{I}) \\ P(\mathbf{w}_{i,j}) &= \mathcal{N}_{\mathbf{w}_{i,j}}(\mathbf{0}, \mathbf{I}) \end{aligned} \tag{4}$$

The PLDA graphical model is illustrated in Figure 6 and the implementation used for this follows the work in (Jiang et al., 2012; Lee et al., 2013a).

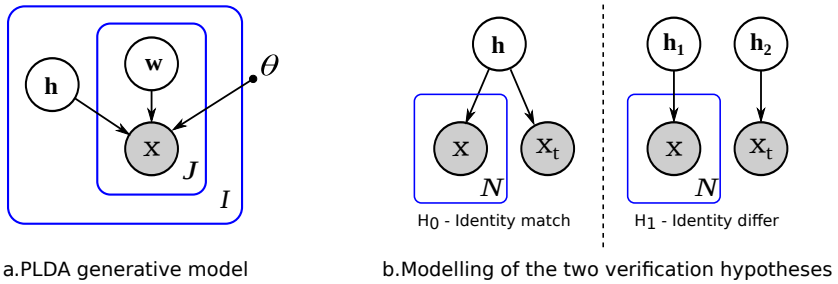


Figure 6: a. Graphical model for the PLDA generative model. For each of the I speakers, J i-vectors $\mathbf{x}$ are observed in the Total Variability space. The PLDA model is described by a set of parameters $\theta = \{\boldsymbol{\mu}, \mathbf{F}, \mathbf{G}, \boldsymbol{\Sigma}\}$
b. Graphical model of the two verification hypotheses considered in the native PLDA framework. The NULL hypothesis, H_0, considers that the N enrollment i-vectors $\mathbf{x}$ and the the test i-vector $\mathbf{x}_t$ belongs to the same speaker and have the same latent variable $\mathbf{h}$ when the alternative hypothesis, H_1, considers that they belong to different speakers and have separate latent variables $\mathbf{h}_1$ and $\mathbf{h}_2$.

4.3.3. I-vector configuration

For this work, the configuration of the i-vector has been chosen empirically to optimize the performance on the development data. A gender-independent 2048-distribution UBM with diagonal covariance matrix is trained on 12,706 sessions from NIST-SRE 2004, 2005 and 2006. A gender-independent Total Variability matrix of rank 400 is then trained by using 10 iterations of EM algorithm described in (Kenny and Dumouchel, 2004) on 66,702 sessions from SwitchBoard II Phase 2 and 3, SwitchBoard Cellular, Part I and II, Fisher English and NIST-SRE 2004, 2005 and 2006 databases. A gender-independent PLDA model is estimated on 26,136 sessions from the 50 male and 47 female speakers of the *background* set of *RSR2015* database. The rank of the matrix $\mathbf{F}$ is set to 400, the matrix $\mathbf{G}$ is set to zero and $\mathbf{\Sigma}$ is full.

In our previous work (Larcher et al., 2013c) we found that using an appropriate definition of the classes used to train the *SphNorm* and PLDA improves the performance of the i-vector system for the case of text-dependent speaker verification. Thus, for experiments on Part I and II, the classes are defined by considering both speaker identity and lexical content of the utterances when they are trained per speaker for experiments on Part III. All component of the i-vector system have been implemented using the open-source toolkit ALIZE (Larcher et al., 2013a).

5. Experimental protocols and results

The rest of this section describes the experimental protocols proposed for the three parts of the *RSR2015* database and performance of the two systems. We will discuss the experiments on the three parts sepaately. To begin with, let's highlight a number of common characteristics to allow for comparison of the systems across the different tasks. In order to develop an evaluation framework, the 300 speakers of the *RSR2015* database are divided into three groups referred to as *background*, *development* and *evaluation*. Although different settings are possible, we propose here a reference protocol that aims at promoting the comparison of algorithms for text-dependent speaker verification. Recordings from the *background* speakers can be used for any purpose, including estimation of the meta-parameters of the speaker verification systems. Decision threshold and possible calibration parameters can be estimated on the *development* part as the *evaluation* set is used for validation. Partitioning of the speakers is given in Table 2.

Table 2: Partitioning of male and female speakers into three groups consisting of *background*, *development* and *evaluation* sets.

Set	**Number of speakers**	
	Male	**Female**
Background	50	47
Development	50	47
Evaluation	57	49

All trials are gender dependent and involve speakers within the same set (*development* or *evaluation*). As described in Section 3.2, each speaker was given a set of three portable devices - A,B,C - to record the nine sessions following the sequence: $\{A,B,C,A,B,C,A,B,C\}$. In order to maximize the mismatch between enrollment and test, sessions {1,4,7}, recorded on device A, are used for enrollment while sessions {2,3,5,6,8,9}, recorded on devices B and C, are used for test[9]. However, a limited inter-session variability might be captured during the modeling as the enrollment utterances come from three different sessions. Multiple models trained per speaker are tested against all test utterances from the other speakers of the same set and gender. The number of trials generated for each part of the database and gender is given in the corresponding sections. All protocols are designed so that the speaker enrollment duration is around 3×3 seconds (3 utterances per enrollment) as this limitation seems reasonable for a commercial application.

5.1. Experimental setup

All systems use the same front-end processing. The training of a state-of-the-art i-vector extractor requires a large amount of data which is only available in 8 kHz telephone channel. For this reason, all data used in this work have been made compatible with our development data by down-sampling the signal to 8 kHz. A bandpass filter (300-3,400 Hz) has then been applied for compatibility with the telephone channel.

Spro[10] is used to extract 19 Mel-Frequency Cepstral Coefficients (MFCC) and the log-energy on a 20 ms sliding window with a shifting of 10 ms between two frames. The first derivatives as well as eleven second derivatives are added to form a feature vector of dimension 50. The normalized log-energy

[9]Note that devices A,B,C can be different between speakers.

[10]http://www.irisa.fr/metiss/guig/spro/

(zero mean, unique variance), is used to select high energy frames based on a two Gaussian distribution model trained for each speech segment. Mean and Variance normalization (MVN) is then applied to the remaining frames.

5.2. Performance measure

Text-independent speaker verification only considers two classes of trials whether the speaker who produces the test utterance is the target speaker or not. Text-dependent speaker verification can be seen as a classification task involving four types of trials whether the speaker who produces the test utterance is the target speaker or not and whether the test-utterance matches the lexical constraint or not (Table 3). Out of these four types of trials, the case where the target speaker pronounces the correct lexical content is regarded as target trial while the three other types of trials should be considered as non-target. Indeed, an impostor should be rejected regardless of the lexical content that (s)he pronounces. Note that the case where the impostor pronounces the correct lexical content (*IMP-correct*) is a genuine imposture that is likely to be more difficult to reject than a naive impostor pronouncing a different lexical content (*IMP-wrong*). Additionally, the case where test utterance is pronounced by the target speaker but does not match the lexical content (*TAR-wrong*) should be rejected as it could be an impostor playing back a recording from the target speaker.

Table 3: The different types of trials defined for text-dependent speaker verification.

	Correct lexical content	Wrong lexical content
Target	*TAR-correct*	*TAR-wrong*
Impostor	*IMP-correct*	*IMP-wrong*

The cost of accepting any of the three types of non-target trials depends of the application so as the probability of each type of trial depends on the deployment conditions. Thus, in order to allow a fair comparison of the systems, performance will be presented for each type of non-target trials separately in terms of equal error rate (EER) and minimum cost ($\underset{\theta}{\operatorname{argmin}}\, C_{DET}(\theta)$) by considering the decision cost function (DCF) given by:

$$C_{DET}(\theta) = C_{Miss} \times P_{Miss}(\theta) \times P_{Target} + C_{FA} \times P_{FA}(\theta) \times (1 - P_{Target}) \quad (5)$$

where C_{Miss} and C_{FA} are the relative costs of detection errors, P_{Miss} and P_{FA} are the miss and false alarm error probabilities and P_{Target} is the a

priori probability of a target speaker. The values for the different parameters are those used for the NIST Speaker Recognition evaluation until 2008 (Przybocki et al., 2006); i.e. $(C_{Miss}, C_{FA}, P_{Target}) = (10, 1, 0.01)$

5.3. Experiments on Part I and II

Due to the similar structure shared by these two parts, a unique protocol is defined to allow an easier comparison. Part I and II address similar scenarios where each speaker pronounces his own pass-phrase, chosen or generated by the system. For each session, the speakers pronounce 30 short sentences in Part I while they pronounce 30 commands in Part II. Part I and II mainly differ in two points. First, utterances from Part II have an average nominal speech duration which is half of the average of Part I (0.63 s against 1.25 s, see Section 3.3). Second, Part II is designed for the task of user-loaded command control in which lexical content of different commands strongly overlap, e.g., *"Volume up"* and *"Volume down"*. Thus Part II is expected to be more difficult than Part I.

5.3.1. Protocol

On Part I, during the enrollment, one model is trained for each of the 30 sentences of a target speaker. The enrollment duration is kept below 10 seconds by using only the three occurrences of this sentence recorded during the enrollment sessions. During the test, the other six occurrences of the same sentence, pronounced by the target speaker in the test sessions, are used to generate *TAR-correct* trials. The other 29 sentences from the 6 test sessions of the target speaker are used to generate *TAR-wrong* trials. Data from all the other speakers from the same set (*development* or *evaluation*) are used to generate impostor trials. The same protocol is applied for the Part II.

5.3.2. Results on Part I

Tables 5 and 6 summarize the performance of the HiLAM and the *i*-vector system on the Part I of the *RSR2015* database for the *development* and *evaluation* sets respectively. The number of trials for each test set is given per gender in Table 4.

Table 4: Number of trials performed on the Part I of the *RSR2015* database for each of the four classes defined for text-dependent speaker verification. The number of trials is given for both male and female protocols on *development* and *evaluation* sets.

Speaker	Lexical content	Male		Female	
		development	*evaluation*	*development*	*evaluation*
TAR	*correct*	8,931	10,244	8,419	8,631
TAR	*wrong*	259,001	297,076	244,123	250,299
IMP	*correct*	437,631	573,664	387,230	414,249
IMP	*wrong*	6,342,019	8,318,132	5,612,176	6,006,596

The nomenclature is as follows: *TAR* refers to the target speaker. *IMP* refers to an impostor speaker. A *correct* lexical content means that the test utterance exactly matches the training material. A *wrong* lexical content means that training and test utterances are different.

The HiLAM system, based on GMM and HMM, outperforms the i-vector system for all definitions of non-target trials, regardless of the speaker's gender and the test set (Tables 5 and 6). The EER obtained by the HiLAM system is at most 66% of the one obtained by the i-vector system (male *development* set considering *IMP-wrong* trials) while, in the best case (female *evaluation* set considering *IMP-wrong* trials) the EER of the HiLAM system is only 18% of the value obtained by the i-vector system. The better performance of the HiLAM system was expected due to the short duration of the training and test utterances as well as the limited channel variability of the dataset (Stafylakis et al., 2013). Additionally, it can be observed on Figure 7-a that, for the *evaluation* male set, the advantage of the HiLAM over the i-vector system persists though all operating regions of the DET curve. Similar behavior has been observed for other sub-sets.

Comparing the performance across genders, performance of the i-vector system is consistent with observations reported in the context of the NIST-SRE *evaluation* where error rates are usually lower or equivalent for the male speakers. Error rates of the HiLAM system are however lower for the female speakers for two of the three definitions of the non-target trials on the *development* set and on the *evaluation* set. A possible explanation for this phenomenon may be the different repartition of speaker specific and lexical information in the frequency bands. A preliminary analysis suggests that a large part of the speaker specific information, located in high frequency for the female is discarded when down-sampling to 8 kHz while more information remains for the male speakers. For this reason, the influence of lexical

Table 5: Performance of *HiLAM* and *i*-vector systems on the *development* set of Part I in terms of Equal Error Rate and minimum DCF (EER % / minDCF×100) for different definitions of target and non-target trials.

User	Target		Impostor		Male		Female	
Text	correct	wrong	correct	wrong	HiLAM	*i*-vector	HiLAM	*i*-vector
Trials	tar	non	-	-	1.66 / 7.40	2.87 / 13.56	1.77 / 7.42	3.05 / 17.26
	tar	-	non	-	3.69 / 16.78	5.95 / 26.74	3.24 / 15.39	7.87 / 40.45
	tar	-	-	non	0.49 / 1.65	0.74 / 3.43	0.45 / 1.81	0.94 / 4.65

The nomenclature is as follows: a *correct* text means that the test utterance exactly matches the training material; a *wrong* text means that training and test utterances are different.

Table 6: Performance of *HiLAM* and *i*-vector systems on the *evaluation* set of Part I in terms of Equal Error Rate and minimum DCF (EER % / minDCF×100) for different definitions of target and non-target trials.

User	Target		Impostor		Male		Female	
Text	correct	wrong	correct	wrong	HiLAM	*i*-vector	HiLAM	*i*-vector
Trials	tar	non	-	-	0.82 / 4.62	1.95 / 11.83	0.61 / 3.44	1.91 / 10.63
	tar	-	non	-	2.47 / 13.51	4.03 / 21.39	2.96 / 15.58	6.61 / 32.69
	tar	-	-	non	0.19 / 0.87	0.32 / 1.88	0.14 / 0.80	0.75 / 3.56

The nomenclature is as follows: a *correct* text means that the test utterance exactly matches the training material; a *wrong* text means that training and test utterances are different.

information may affect more the female speakers than the male.

Results reported in Tables 5 and 6 show that the HiLAM system, modeling each sentence by a 5-state HMM can better reject the target speaker pronouncing a wrong sentence than an impostor who knows the correct pass-phrase. This result shows the efficiency of the HMM to model the temporal structure of the pass-phrase, even with a limited amount of training data. The same conclusion stands for the *i*-vector system that does not model any temporal information but includes lexical information through the *i*-vector normalization and PLDA training (Larcher et al., 2013c). Indeed, for all male and female trials on both *development* and *evaluation* sets, the lexical information conveyed by the *i*-vectors seem predominant compared to the speaker information as shown in (Larcher et al., 2012a).

5.3.3. Results on Part II

Tables 8 and 9 summarize the performance of the HiLAM and the *i*-vector system on Part II of the *RSR2015* database for the *development* and *evaluation* sets respectively. The number of trials for each test set is given per gender in Table 7.

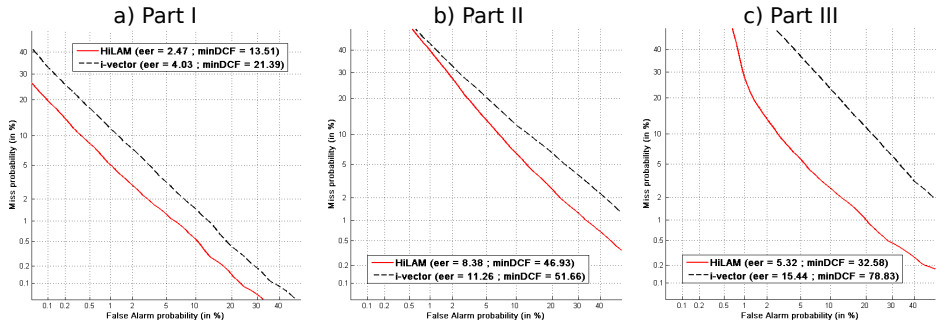


Figure 7: Detection error trade-off (DET) curves for the male *evaluation* sets of Part I, II and III of the *RSR2015* database. In all trials, target and impostor speaker pronounce the *correct* text, i.e., the test utterance exactly matches the training material or the prompted sequence of digits.

Table 7: Number of trials performed on the Part II of the *RSR2015* database for each of the four classes defined for text-dependent speaker verification. The number of trials is given for both male and female protocols on *development* and *evaluation* sets.

Speaker	Lexical content	Male		Female	
		development	*evaluation*	*development*	*evaluation*
TAR	*correct*	8,960	10,238	8,444	8,637
TAR	*wrong*	259,841	296,902	244,876	250,473
IMP	*correct*	439,042	573,328	388,424	414,579
IMP	*wrong*	6,361,855	8,311,644	5,630,820	6,009,351

The nomenclature is as follows: *TAR* refers to the target speaker. *IMP* refers to an impostor speaker. A *correct* lexical content means that the test utterance exactly matches the training material. A *wrong* lexical content means that training and test utterances are different.

As expected, both systems suffer from the short duration and the lexical similarity of the commands from the Part II of the *RSR2015* database. Compared to Part I where sentences are twice longer in average, the EERs on Part II are at least 61 % higher than on Part I for the same set of speakers (female *development* set when considering *IMP-correct* trials). In the worst case, for the female speakers of the *evaluation* set when considering *IMP-wrong* trials, the error rate increases by 903%.

For the female speakers, and similarly to the experiments on Part I, an important performance gap in favor of HiLAM system can be observed in all configurations. On the opposite, the gap between the two systems is

Table 8: Performance of *HiLAM* and *i*-vector systems on the *development* set of Part II in terms of Equal Error Rate and minimum DCF (EER % / minDCF×100) for different definitions of target and non-target access.

User	Target		Impostor		Male		Female	
Text	correct	wrong	correct	wrong	HiLAM	*i*-vector	HiLAM	*i*-vector
Trials	tar	non	-	-	6.14 / 34.40	5.41 / 32.19	4.62 / 28.16	6.94 / 43.04
	tar	-	non	-	10.58 / 50.24	13.75 / 58.34	6.66 / 30.10	12.73 / 57.62
	tar	-	-	non	3.03 / 13.36	2.50 / 12.68	1.29 / 5.94	2.86 / 14.26

The nomenclature is as follows: a *correct* text means that the test utterance exactly matches the training material; a *wrong* text means that training and test utterances are different.

Table 9: Performance of *HiLAM* and *i*-vector systems on the *evaluation* set of Part II in terms of Equal Error Rate and minimum DCF (EER % / minDCF×100) for different definitions of target and non-target access.

User	Target		Impostor		Male		Female	
Text	correct	wrong	correct	wrong	HiLAM	*i*-vector	HiLAM	*i*-vector
Trials	tar	non	-	-	4.42 / 28.05	4.39 / 26.66	3.71 / 23.48	5.16 / 28.79
	tar	-	non	-	8.38 / 46.93	11.26 / 51.66	7.95 / 40.04	15.27 / 67.01
	tar	-	-	non	1.71 / 8.80	1.81 / 9.59	1.45 / 7.14	3.05 / 15.06

The nomenclature is as follows: a *correct* text means that the test utterance exactly matches the training material; a *wrong* text means that training and test utterances are different.

reduced for the male speakers. For non-target trials where the target speaker pronounces a wrong lexical content, the best performance is even obtained by the *i*-vector system on both *development* and *evaluation* sets as well as for the case of impostor pronouncing a wrong lexical content in *development* set. The curved DET plot obtained for the HiLAM system on Figure 7-b shows that the score distributions of the HiLAM system are less Gaussian than for the Part I while the *i*-vector does not suffer from such effect.

Performance across genders follows the same trend as for the Part I. In all configurations except impostor pronouncing the correct command for the *development* set, the *i*-vector system performs better on male speakers. On the contrary, the HiLAM system consistently performs better for female speakers.

5.4. Part III

Part III of the *RSR2015* database focuses on scenario where the training and test utterances share the same phonetic content but with different context. Thus, co-articulation is different between train and test. Due to the limited lexicon used in this part - only English digits - the UBM of the

HiLAM system is trained only on digit strings from the *background* set speakers. The i-vector system is the same as for the other parts as the quantity of data available is not enough to train the i-vector extractor on digits only.

5.4.1. Protocol

Part III of the *RSR2015* database is designed to evaluate the ability of a system to take advantage of the temporal structure of the prompted sequence. During the test, the sequence of digits for the speaker to pronounce is assumed to be randomly generated. For the HiLAM system, especially modified for this task, the model used for a test is created on the fly to match the prompted sequence. Because the i-vector system does not model the temporal structure of the utterances, the model used for testing is unchanged whatever the prompted digit sequence. Yet, the i-vector system only considers two types of trials, *TAR-correct* and *IMP-correct*, as the lexical content is not taken into account any more. Performance of the i-vector system, is given to evaluate the degradation caused by the mismatch of co-articulation.

For the HiLAM system, during the enrollment, one set of digit models is trained for each enrollment session of a target speaker. Using only the three ten-digit sequences pronounced by the target speaker in this session keeps the enrollment duration around 15 seconds. During the test, all five-digit sequence's prompts from the six test sessions are used to generate trials. Duration of the test utterance is thus comparable with Part I. For each of those 60 prompts, the specific model created by the HiLAM is compared to all five-digit recordings from all speakers of the test set.

Four types of trials are defined whether the speaker is the target speaker (TAR) or an impostor (IMP) and whether the lexical content, i.e., the digit sequence, matches the sequence prompted by the system at test time (*correct*) or is different (*wrong*). Note that the definition of trials involving *correct* lexical content is different from the one given in Part I and II. A *correct* lexical content in Part I and II was defined according to the training utterances while in Part III it is define according to the prompted utterance. For any trial, the model created at test time by the HiLAM system exactly match the prompted digit sequence. The number of trials of each category are given in Table 10.

5.4.2. Results on Part III

Tables 11 and 12 summarize the performance of the HiLAM and the i-vector system on the Part III of the *RSR2015* database for the *development*

Table 10: Number of trials performed on the Part III of the *RSR2015* database for each of the four classes defined for text-dependent speaker verification. The number of trials is given for both male and female protocols on *development* and *evaluation* sets.

Speaker	Lexical content	Male		Female	
		development	*evaluation*	*development*	*evaluation*
TAR	*correct*	5,154	5,943	5,025	5,283
TAR	*wrong*	412,968	476,331	402,405	422,883
IMP	*correct*	251,310	332,863	231,155	253,584
IMP	*wrong*	10,022,832	13,255,958	9,197,556	10,085,760

The nomenclature is as follows: *TAR* refers to the target speaker. *IMP* refers to an impostor speaker. A *correct* lexical content means that the test utterance exactly matches the prompted sequence of digits. A *wrong* lexical content means that the sequence of digits pronounced during the test is different form the prompted one.

and *evaluation* sets.

First it can be noticed that the HiLAM system does not reach the same performance as when compared to Part I despite comparable test durations. Our results are consistent with the work in (Kato and Shimizu, 2003; Hébert, 2008) when lexical content is not kept. The authors of (Kato and Shimizu, 2003) report that "preserving digit strings improves accuracy" by a relative 50%. The benefit of co-articulation is even higher for the HiLAM system as EER on Part III is higher by 76% relative for male and by 227% for female on *development* set (115% and 267% for *evaluation* set) when compared to Part I in *IMP-correct* non-target definition. Second, performance on female speakers are significantly worse than the one on male speakers for both systems and test sets.

Performance of the HiLAM system is very poor when discriminating between *correct* and *wrong* lexical content (line 3 of Tables 11 and 12). This may be due to the modeling of each digit by a single state and to the adaptation of this state from the digit-independent GMM from the second layer of the architecture. Modeling each digit by several states may improve the performance of the system as the mismatch of the co-articulation would not affect the whole digit model. The same conclusion stands when comparing the performance between *IMP-correct* and *IMP-wrong*. The influence of the lexical mismatch to help the verification system to reject *IMP-wrong* is not as important as for Part I or II. On Figure 7-c, we observe that the higher

Table 11: Performance of *HiLAM* system on the *development* set of Part III in terms of Equal Error Rate and minimum DCF (EER % / minDCF×100) for different definitions of target and non-target access.

User	Target		Impostor		Male		Female	
Text	correct	wrong	correct	wrong	HiLAM	*i*-vector	HiLAM	*i*-vector
Trials	tar	non	-	-	38.32 / 99.96		38.35 / 98.18	
	tar	-	non	-	6.50 / 33.39	16.37 / 69.09	10.60 / 44.30	18.56 / 80.78
	tar	-	-	non	6.13 / 29.84		10.55 / 40.00	

The nomenclature is as follows: a *correct* text means that the test utterance exactly matches the prompted sequence of digits; a *wrong* text means that the sequence of digits pronounced during the test is different form the prompted one.

Table 12: Performance of *HiLAM* system on the *evaluation* set of Part III in terms of Equal Error Rate and minimum DCF (EER % / minDCF×100) for different definitions of target and non-target access.

User	Target		Impostor		Male		Female	
Text	correct	wrong	correct	wrong	HiLAM	*i*-vector	HiLAM	*i*-vector
Trials	tar	non	-	-	36.41 / 99.98		38.78 / 98.31	
	tar	-	non	-	5.32 / 32.58	15.44 / 73.83	10.87 / 46.86	25.63 / 93.67
	tar	-	-	non	4.88 / 27.95		10.07 / 40.10	

The nomenclature is as follows: a *correct* text means that the test utterance exactly matches prompted sequence of digits; *wrong* text means that the sequence of digits pronounced during the test is different form the prompted one.

part of the DET plot obtained for the HiLAM is strongly curved due to non-Gaussian score distributions. Again, the DET plot of the *i*-vector / PLDA system is straight, confirming that this system generate more Gaussian score distributions.

Results of the *i*-vector system on Part III can be compared to condition *IMP-correct* of Part I as the phonetic content conveyed by the *i*-vector from the test utterance is a subset of the phonetic content from the enrollment material. Nonetheless, temporal structure of enrollment and test is not exactly matching in Part III while the temporal structure of enrollment and test exactly matches in Part I. Results of the *i*-vector system confirm the importance of matching the exact lexical content, including co-articulation. Indeed, EERs on Part III are at least 76% higher than for the condition *IMP-correct* of Part I (Tables 5 and 6).

6. Distribution

The *RSR2015* database is distributed at a nominal cost in order to support the continuous effort of text-dependent speaker verification database

development. The main goal of the distribution is to provide a framework for comparison of algorithms and systems across the community. Institutions willing to acquire the database will have to sign a license agreement that has been made available on ETPL website[11] since 2012. ETPL is the technology transfer arm of the Agency for Science, Technology and Research (A*STAR) in Singapore.

7. Conclusion

Among the three contributions presented in this paper, the main one is the release of a large corpus, the *RSR2015* database. The *RSR2015* database, has been collected and made available with the aim of allowing comparison of text-dependent speaker verification algorithms under different duration and lexical constraints. As all speakers repeat the same pass-phrases in different sessions, we believe that the *RSR2015* database can also be used to conduct anti-playback analysis. The *RSR2015* database includes 151 hours of speech signal recorded from 300 gender-balanced speakers and is one of the largest corpus publicly available for text-dependent speaker verification.

As a second contribution, we produced the largest inventory of speech databases for text-dependent task available in the literature to our best knowledge. We presented the tendencies and main characteristics of existing databases that led to the design of the *RSR2015* database. Despite the huge effort of the community to produce large and usable databases in the recent years, we highlighted several lacks in the existing databases. The necessary large number of speakers, the need of a balanced gender representation and the duration and lexical variability motivated the collection of the *RSR2015* database. Recent publications applying resource intensive methods developed for text-independent task testify of the contribution of the *RSR2015* database to fill the gap between text-dependent and text-independent research fields (Larcher et al., 2012a, 2013c; Stafylakis et al., 2013). Together with this survey of databases, we produced a description of existing classifiers dedicated to text-dependent speaker verification.

The third contribution of this paper consists of evaluation protocols proposed for each of the three parts of the *RSR2015* database. The protocols

[11] `http://www.etpl.sg/innovation-offerings/ready-to-sign-licenses/rsr2015-overview-n-specifications` (Accessed February 28, 2014)

allow comparison of algorithms in the different tasks covered by the *RSR2015* database. Performance of two systems are given as a baseline and compared on the different protocols, the HiLAM text-dependent system based on GMM and HMM modeling (Larcher et al., 2012b; Lee et al., 2011) and a state-of-the-art i-vector/PLDA system based on the open source ALIZE toolkit (Larcher et al., 2013a).

Experiments show that our GMM/HMM-based system outperforms the i-vector system in most of the configurations. This confirms the well known weakness of i-vector systems on short durations that has recently been widely studied (Kenny et al., 2013; Cumani et al., 2013; Hasan et al., 2013). Behavior of the i-vector system in the context of short duration text-dependent speaker verification is consistent with the previous studies in the context of text-independent speaker verification (Senoussaoui et al., 2011), reaching lower error rates on male speakers. On the opposite, the HiLAM system performs better on female speakers on both Part I and II of the *RSR2015* database. This behavior will be the focus of future work investigating the distribution of speaker and lexical information in the speech signal.

For the case of fixed pass-phrases (Part I and II), we found that it is easier to reject an attack where the impostor plays back a recording of the target speaker pronouncing a text-different from the expected pass-phrase than an impostor pronouncing the correct pass-phrase. This confirms observations from (Larcher et al., 2012a) that lexical information is dominating in short speech segments, even for the case of the i-vector system, despite the lack of consideration for the temporal structure of the utterances. Performances of both systems are strongly affected by the co-articulation mismatch inherent to the randomly prompted digit scenario (Part III of the *RSR2015* database). Compared to Part I which offers similar speech durations, degradations caused by co-articulation mismatch are found to be equivalent or higher than the one reported in (Hébert, 2008; Kato and Shimizu, 2003). The increase of error rates observed for the i-vector system shows that methods, without exploiting the temporal information of the speech signal, suffer from the co-articulation effect.

An extension of the *RSR2015* database is being recorded to include more challenging recording conditions. This part consists of the Part I being transmitted over marine VHF channel.

Appendix A. The RSR2015 lexical content

Table A.13: List of the prompts recorded by all speakers for each of the 9 sessions of the *RSR2015* database.

Part I	**Part II**	**Part III**
Only lawyers love millionaires	Watch movie	1 - 7 - 4 - 0 - 9 - 3 - 8 - 2 - 5 - 6
No return address whatsoever	Watch cartoon	3 - 7 - 0 - 8 - 6 - 9 - 5 - 1 - 4 - 2
Do without fancy tablecloths	Play music	8 - 1 - 5 - 9 - 0 - 6 - 7 - 4 - 2 - 3
She can remove all knick knacks within reach	Play Game	4 - 8 - 0 - 7 - 3
I know I didn't meet her early enough	Call mum	1 - 9 - 6 - 5 - 2
Artificial intelligence is for real	Call dad	9 - 3 - 6 - 0 - 1
Allow each child to have an ice pop	Call sister	2 - 7 - 4 - 8 - 5
When she awoke she was the ship	Call brother	7 - 2 - 9 - 0 - 5
Well now we have two big theaters	Coffee	6 - 3 - 1 - 4 - 8
Toss a die until an ace appears	Cappuccino	8 - 6 - 2 - 3 - 9
This coat looks like a rag heap	Espresso	5 - 4 - 0 - 7 - 1
My dress needs some work on it	Door open	0 - 6 - 4 - 9 - 2
It was time to go up myself	Door close	5 - 8 - 7 - 3 - 1
He would not carry a brief case	Door hold	
He felt a good deal less shaky	Turn on Master	
Do buy all purpose mugs or cups	Master off	
By eating yogurt you may live longer	Turn on TV	
But how little love we give him	TV off	
Yet we no longer feel uneasy	Turn on dish washer	
She is thinner than I am	Dish washer off	
The drunkard is a social outcast	Turn on coffee machine	
The Birthday party has cupcake and ice-cream	Coffee machine off	
A good attitude is unbeatable	Turn light on	
Basketball can be an entertaining sport	Light off	
And so he walked aimless again	Turn on air-con	
A huge power outage rarely occurs	Air-con off	
Guerrillas were racing toward him	Turn on oven	
The rose corsage smelled sweet	Oven off	
There was typhoid and malaria	Volume up	
The redcoats ran like rabbits	Volume down	

Speech material for Part I and II is fixed across the 9 sessions while digit sequences in Part III vary from session to session but are kept the same across speakers. These digit sequences are given as an example as the sequences vary across sessions. A complete transcription of the speech material recorded for the Part III is provided with the database.

References

Amino, K., Arai, T., 2009. Speaker-dependent characteristics of the nasals. Forensic science international 185, 21–28.

Aronowitz, H., 2012. Text-Dependent Speaker Verification Using a Small Development Set, in: Odyssey Speaker and Language Recognition Workshop.

Avinash, B., Guruprasad, S., Ygnannarayana, B., 2010. Exploring subsegmental and suprasegmental features for a text-dependent speaker verification in distant speech signals, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 1073–1076.

Bailly-Bailliere, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Marietho z, J., Matas, J., Messer, K., Popovici, V., Poree, F., et al., 2003. The BANCA database and evaluation protocol. Lecture Notes in Computer Science (LNCS) 2688, 625–638.

BenZeghiba, M.F., Bourlard, H., 2006. User-customized password speaker verification using multiple reference and background models. Speech Communication 48, 1200–1213.

Boakye, K., Peskin, B., 2004. Text-Constrained Speaker Recognition on a Text-Independent Task, in: Odyssey Speaker and Language Recognition Workshop, pp. 1–6.

Boies, D., Hébert, M., Heck, L.P., 2004. Study on the effect of lexical mismatch in text-dependent speaker verification, in: Odyssey Speaker and Language Recognition Workshop, pp. 1–5.

Bonastre, J.F., Morin, P., Junqua, J.C., 2003. Gaussian dynamic warping (GDW) method applied to text-dependent speaker detection and verification, in: European Conference on Speech Communication and Technology (Eurospeech), pp. 2013–2016.

Bousquet, P.M., Larcher, A., Matrouf, D., Bonastre, J.F., Plchot, O., 2012. Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis, in: Odyssey Speaker and Language Recognition Workshop, pp. 1–8.

Bousquet, P.M., Matrouf, D., Bonastre, J.F., 2011. Intersession compensation and scoring methods in the i-vectors space for speaker recognition, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 485–488.

Brümmer, N., de Villiers, E., 2010. The speaker partitioning problem, in: Odyssey Speaker and Language Recognition Workshop, pp. 1–8.

Campbell, J., Higgins, A.L., 1994. A YOHO speaker verification corpus LDC94s16 (available on LCD website: http://www.ldc.upenn.edu).

Campbell, J.P., 1995. Testing with the YOHO CD-ROM voice verification corpus, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 341–344.

Campbell, J.P., Reynolds, D.A., 1999. Corpora for the evaluation of speaker recognition systems, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 829–832.

Campbell, J.P., Shen, W., Campbell, W.M., Schwartz, R., Bonastre, J.F., Matrouf, D., 2009. Forensic speaker recognition. Signal processing magazine, IEEE 26, 95–103.

Charlet, D., Jouvet, D., 1997. Optimizing feature set for speaker verification. Pattern Recognition Letters 18, 873–879.

Charlet, D., Jouvet, D., Collin, O., 2000. An alternative normalization scheme in HMM-based text-dependent speaker verification. Speech Communication 31, 113–120.

Chatzis, S., Varvarigou, T., 2007. A Robust to Outliers Hidden Markov Model with Application in Text-Dependent Speaker Identification, in: International Conference on Signal Processing and Communications, pp. 804–807.

Che, C.W., Lin, Q., Yuk, D.S., 1996. An HMM approach to text-prompted speaker verification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 673–676.

Chen, K., Xie, D., Chi, H., 1996. A modified HME architecture for text-dependent speaker identification. IEEE Transactions on Neural Networks 7, 1309–1313. doi:10.1109/72.536325.

Chen, W., Hong, Q., Li, X., 2012. GMM-UBM for text-dependent speaker recognition, in: International Conference on Audio, Language and Image Processing (ICALIP), IEEE. pp. 432–435.

Chollet, G., Cochard, J.L., Constantinescu, A., Jaboulet, C., Langlais, P., 1996. Swiss French PolyPhone and PolyVar: telephone speech databases to model inter- and intra-speaker variability. Technical Report. IDIAP.

Cole, R., Noel, M., Noel, V., 1998. The CSLU speaker recognition corpus, in: Proceedings International Conference on Spoken Language Processing, ICSLP, pp. 3167–3170.

Cumani, S., Glembek, O., Brummer, N., de Villiers, E., Laface, P., 2012. Gender independent discriminative speaker recognition in i-vector space, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 4361–4364.

Cumani, S., Plchot, O., Laface, P., 2013. Probabilistic Linear Discriminant Analysis of I-Vector Posterior Distribution, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 7644–7647.

Das, A., Tapaswi, M., 2010. Direct modeling of spoken passwords for text-dependent speaker recognition by compressed time-feature representations, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 4510–4513.

Dehak, N., Dehak, R., Glass, J., Reynolds, D., Kenny, P., 2010. Cosine similarity scoring without score normalization techniques, in: Odyssey Speaker and Language Recognition Workshop, Odyssey. pp. 1–5.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011a. Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing 19, 788–798.

Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D.A., Dehak, R., 2011b. Language Recognition via i-vectors and Dimensionality Reduction, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 857–860.

Dessimoz, D., Richiardi, J., Champod, C., Drygajlo, A., 2008. Multimodal biometrics for identity documents. Forensic science international 167, 154–159.

Dialogues Spotlight Technology, 2000. Large Scale Evaluation of Automatic Speaker Verification Technology. Technical Report. The Center for Communication Interface Research, University of Edinburgh.

Doddington, G., 2012. The Effect of Target/Non-Target Age Difference on Speaker Recognition Performance, in: Odyssey Speaker and Language Recognition Workshop, pp. 1–5.

Doddington, G.R., 1998. Speaker recognition evaluation methodology–an overview and perspective–, in: Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pp. 20–23.

Dong, C., Dong, Y., Li, J., Wang, H., 2008. Support Vector Machines Based Text Dependent Speaker Verification Using HMM superverctors, in: Odyssey Speaker and Language Recognition Workshop, pp. 1–7.

Dumas, B., Pugin, C., Hennebert, J., Petrovska-Delacrétaz, D., Humm, A., Evéquoz, F., Ingold, R., Rotz, D.V., 2005. MyIdea–Multimodal biometrics database, description of acquisition protocols. Biometrics on the Internet 275, 59–62.

Dutta, T., 2007. Text dependent speaker identification based on spectrograms, in: Image and Vision Computing, pp. 238–243.

Dutta, T., 2008. Dynamic Time Warping Based Approach to Text-Dependent Speaker Identification Using Spectrograms, in: Congress on Image and Signal Processing, pp. 354–360.

ELDA - Evaluations and Language resources Distribution Agency, 2003. S0050, RUSTEN: Russian Switched Telephone Network speech database (STC). URL: `http://www.elda.fr/catalogue/en/speech/S0050.html`.

Farrell, K.R., 1995. Text-dependent speaker verification using data fusion, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, INSTITUTE OF ELECTRICAL ENGINEERS INC (IEE). pp. 349–349.

Farrell, K.R., Ramachandran, R.P., Mammone, R.J., 1998. An analysis of data fusion methods for speaker verification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 1129–1132. doi:`10.1109/ICASSP.1998.675468`.

Faundez-Zanuy, M., Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., 2006. Multimodal biometric databases: An overview. IEEE Aerospace and Electronic Systems Magazine 21, 29–37.

Fauve, B., 2009. Tackling Variabilities in Speaker Verification with a Focus on Short Durations. Ph.D. thesis. School of Engineering Swansea University.

Fierrez, J., Galbally, J., Ortega-Garcia, J., Freire, M., Alonso-Fernandez, F., Ramos, D., Toledano, D., Gonzalez-Rodriguez, J., Siguenza, J., Garrido-Salas, J., et al., 2010. BiosecurID: a multimodal biometric database. Pattern Analysis & Applications 13, 235–246.

Fierrez, J., Ortega-Garcia, J., Torre Toledano, D., Gonzalez-Rodriguez, J., 2007. Biosec baseline corpus: A multimodal biometric database. Pattern Recognition 40, 1389–1392.

Finan, R., Sapeluk, A., Damper, R., 1996. Comparison of multilayer and radial basis function neural networks for text-dependent speaker recognition, in: IEEE International Conference on Neural Networks, IEEE. pp. 1992–1997.

Forsyth, M., 1995. Discriminating observation probability (DOP) HMM for speaker verification. Speech communication 17, 117–129.

Fox, N.A., O'Mullane, B.A., Reilly, R.B., 2005. The Realistic Multi-modal VALID database and Visual Speaker Identification Comparison Experiments, in: International Conference of Audio and Video-Based Person Authentication, AVBPA, New York (US). pp. 777–786.

Furui, S., 1981a. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing] 29, 254–272.

Furui, S., 1981b. Comparison of speaker recognition methods using statistical features and dynamic features. IEEE Transactions on Acoustics, Speech and Signal Processing 29, 342–350.

Garcia-Romero, D., Espy-Wilson, C.Y., 2011. Analysis of i-vector length normalization in speaker recognition systems, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 249–252.

Garcia-Salicetti, S., Beumier, C., Chollet, G., Dorizzi, B., Jardins, J., Lunter, J., Ni, Y., Petrovska-Delacretaz, D., 2003. BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and

Signature Modalities. Lecture Notes in Computer Science 2688/2003, 845–853.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N., Zue, V., 1993. Timit acoustic-phonetic continuous speech corpus linguistic data consortium. Philadelphia, PA 1.

Gu, Y., Thomas, T., 1998. An implementation and evaluation of an on-line speaker verification system for field trials, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 125–128.

Hasan, T., Saeidi, R., Hansen, J.H.L., van Leeuwen, D.A., 2013. Duration Mismatch Compensation for I-Vector Based Speaker Recognition Systems, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 7663–7667.

Hébert, M., 2008. Handbook of Speech Processing. Springer-Verlag, Heidelberg. chapter Text-dependent speaker recognition. pp. 743–762.

Hébert, M., Boies, D., 2005. T-norm for text-dependent commercial speaker verification applications: Effect of lexical mismatch, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 729–732.

Hebert, M., Heck, L.P., 2003. Phonetic class-based speaker verification, in: European Conference on Speech Communication and Technology (Eurospeech), Geneva. pp. 1665–1668.

Heck, L., Genoud, D., 2001. Integrating Speaker and Speech Recognizers: Automatic Identity Claim Capture for Speaker Verification, in: Odyssey Speaker and Language Recognition Workshop, pp. 249–254.

Hennebert, J., Melin, H., Petrovska, D., Genoud, D., 2000. POLYCOST: A telephone-speech database for speaker recognition. Speech Communication 31, 265–270.

Jiang, Y., Lee, K.A., Tang, Z., Ma, B., Larcher, A., Li, H., 2012. PLDA Modeling in I-vector and Supervector Space for Speaker Verification, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 1680–1683.

Kahn, J., Audibert, N., Bonastre, J.F., Rossato, S., 2011. Inter and intra-speaker variability in French: an analysis of oral vowels and itsimplication for automatic speaker verification, in: International Congress of Phonetic Sciences (ICPhS), pp. 1002–1005.

Kahn, J., Audibert, N., Rossato, S., Bonastre, J.F., 2010. Intra-speaker variability effects on Speaker Verification performance, in: Odyssey Speaker and Language Recognition Workshop, pp. 109–116.

Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., Mason, M., 2011. I-vector Based Speaker Recognition on Short Utterances, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 2341–2344.

Karam, Z.N., Campbell, W.M., Dehak, N., 2011. Graph relational features for speaker recognition and mining, in: Statistical Signal Processing Workshop (SSP), IEEE. pp. 525–528.

Karlsson, I., 1999. Within-speaker variability in the VeriVox database. Gothenburg papers in theoretical linguistics , 93–96.

Karlsson, I., Banziger, T., Dankovicová, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K., 2000. Speaker verification with elicited speaking styles in the VeriVox project. Speech Communication 31, 121–129.

Kato, T., Shimizu, T., 2003. Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence-preserving connected digit patterns, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 57–60.

Kekre, H., Sarode, T., Natu, S., Natu, P., 2010. Performance Comparison Of 2-D DCT On Full/Block Spectrogram And 1-D DCT On Row Mean Of Spectrogram For Speaker Identification. International Journal of Biometrics and Bioinformatics (IJBB) 4, 100.

Kelly, F., Drygajlo, A., Harte, N., 2012. Speaker verification with long-term ageing data, in: International Conference on Biometrics (ICB), pp. 478–483.

Kelly, F., Harte, N., 2011. Effects of long-term ageing on speaker verification, in: Biometrics and ID Management. Springer, pp. 113–124.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing 15, 1435–1447.

Kenny, P., Dumouchel, P., 2004. Disentangling speaker and channel effects in speaker verification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 37–40.

Kenny, P., Stafylakis, T., Ouellet, P., Alam, J., Dumouchel, P., 2013. PLDA for Speaker Verification with Utterances of Arbitrary Duration, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 7649–7653.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech Communication 52, 12–40.

Larcher, A., Bonastre, J.F., Fauve, B., Lee, K.A., Lévy, C., Li, H., Mason, J.S., Parfait, J.Y., 2013a. ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 2768–2773.

Larcher, A., Bonastre, J.F., Mason, J.S., 2013b. Reinforced temporal structure of acoustic models for speaker recognition. Digital Signal Processing 23, 1910–1917. URL: http://www.sciencedirect.com/science/article/pii/S1051200413001504, doi:http://dx.doi.org/10.1016/j.dsp.2013.07.007e.

Larcher, A., Bonastre, J.F., Mason, J.S.D., 2008. Reinforced temporal structure information for embedded utterance-based speaker recognition, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 371–374.

Larcher, A., Bousquet, P.M., Lee, K.A., Matrouf, D., Li, H., Bonastre, J.F., 2012a. I-vectors in the context of phonetically-constrained short utterances for speaker verification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 4773–4776.

Larcher, A., Lee, K.A., Ma, B., Li, H., 2012b. The RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 1580–1583.

Larcher, A., Lee, K.A., Ma, B., Li, H., 2013c. Phonetically-Constrained PLDA Modeling for Text-Dependent Speaker Verification with Multiple Short Utterances, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 7673–7677.

Lawson, A.D., Staufer, A., Smolenski, B., Pokines, B., Leoanrd, M., Cupples, E., 2009. Long term examination of intra-session and inter-session speaker variability, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 2899–2902.

Lee, K.A., Larcher, A., Thai, H., Ma, B., Li, H., 2011. Joint Application of Speech and Speaker Recognition for Automation and Security in Smart Home, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 3317–3318.

Lee, K.A., Larcher, A., You, C.H., Ma, B., Li, H., 2013a. Multi-session PLDA Scoring of I-vector for Partially Open-Set Speaker Detection, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 3651–3655.

Lee, K.A., Ma, B., Li, H., 2013b. Speaker verification makes its debut in smartphone, in: SLTC Newsletter.

van Leeuwen, D.A., Brümmer, N., 2013. The distribution of calibrated likelihood-ratios in speaker recognition, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 1619–1623.

Lei, Y., Hansen, J.H., 2009. The Role of Age in Factor Analysis for Speaker Identification, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 2371–2374.

Li, H., Ma, B., Lee, K.A., 2013. Spoken language recognition: from fundamentals to practice. Proceedings of the IEEE 101, 1136–1159.

Li, Q., Zheng, J., Tsai, A., Zhou, Q., 2002. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. IEEE Transactions on Speech and Audio Processing 10, 146–157.

Luan, J., Hao, J., Kakino, T., Ikumi, T., 2006. Template Compression and Distance Normalization for Reliable Text-dependent Speaker Verification, in: Odyssey Speaker and Language Recognition Workshop, IEEE. pp. 1–4.

Mandasari, M.I., McLaren, M., van Leeuwen, D., 2011. Evaluation of i-vector Speaker Recognition Systems for Forensic Application, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 21–24.

Marcel, S., McCool, C., Matejka, P., Cernocky, J., Kittler, J., Glembek, O., Plchot, O., Jancik, Z., Larcher, A., Levy, C., 2010. On the Results of the First Mobile Biometry (MOBIO) Face and Speaker Verification Evaluation. Lecture Notes in Computer Science 2010, 210–225. URL: `http://www.fit.vutbr.cz/research/view_pub.php?id=9449`.

Martin, A.F., Greenberg, C.S., 2009. NIST 2008 speaker recognition evaluation: performance across telephone and room microphone channels, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 2579–2582.

Martin, A.F., Greenberg, C.S., 2010. The NIST 2010 speaker recognition evaluation, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 2726–2729.

Martinez, D., Plchot, O., Burget, L., Glembek, O., Matejka, P., 2011. Language Recognition in i-vectors Space, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 861–864.

Mason, J.S., Deravi, F., Chibelushi, C.C., Gandon, S., 1996. Project: DAVID (Digital Audio Visual Integrated Database). Technical Report. Department of Electrical and Electronic Engineering, University of Wales Swansea.

Matsui, T., Furui, S., 1993. Concatenated phoneme models for text-variable speaker recognition, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 391–394.

Meng, H., Ching, P., Lee, T., Mak, M.W., Mak, B., Moon, Y., Siu, X., Tang, M.H., Tang, X., Hui, H.P., Lee, A., et al., 2006. The multi-biometric, multi-device and multilingual (m3) corpus, in: International Workshop on Multimodal User Authentication, pp. 1–8.

Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G., 1999. XM2VTSDB: The Extended M2VTS Database, in: International Conference of Audio and Video-Based Person Authentication, AVBPA, pp. 965–966.

Mistretta, W., Farrell, K., 1998. Model adaptation methods for speaker verification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, IEEE. pp. 113–116.

Nakagawa, S., Wei, Z., Takahashi, M., 2004. Text-independent speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Montreal (Canada). pp. I–81.

Nosratighods, M., Ambikairajah, E., Epps, J., Carey, M.J., 2010. A segment selection technique for speaker verification. Speech Communication 52, 753–761.

Ortega-Garcia, J., Fierrez, J., Alonso-Fernandez, F., Galbally, J., Freire, M.R., Gonzalez-Rodriguez, J., Garcia-Mateo, C., Alba-Castro, J.L., Gonzalez-Agulla, E., Otero-Muras, E., et al., 2010. The multiscenario multienvironment biosecure multimodal database (bmdb). IEEE transactions on Pattern Analysis and Machine intelligence 32, 1097–1111.

Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V., 2000. AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. Speech communication 31, 255–264.

Pigeon, S., Vandendorpe, L., 1997. The M2VTS multimodal face database (release 1.00). Lecture Notes in Computer Science 1206/1997, 403–409.

Prazak, J., Silovsky, J., 2011. Speaker Diarization Using PLDA-based Speaker Clustering, in: International Conference on Intelligent Data Acquisition and Advanced Computing Systems, pp. 347–350.

Prince, S.J., Elder, J.H., 2007. Probabilistic linear discriminant analysis for inferences about identity, in: International Conference on Computer Vision, IEEE. pp. 1–8.

Przybocki, M.A., Martin, A.F., Le, A.N., 2006. NIST Speaker Recognition Evaluation Chronicles - Part 2, in: Odyssey Speaker and Language Recognition Workshop, pp. 1–6.

Ramasubramanian, V., Das, A., Kumar, V., 2006. Text-Dependent Speaker-Recognition Using One-Pass Dynamic Programming Algorithm, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. I–1.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41. doi:10.1006/dspr.1999.0361.

Rosenberg, A.E., Lee, C., Gokcen, S., 1991. Connected word talker verification using whole word Hidden Markov Models, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 381–384.

Rosenberg, A.E., Siohan, O., Parthasarathy, S., 2000. Small group speaker identification with common password phrases. Speech communication 31, 131–140.

Schmidt, M., Gish, H., 1996. Speaker identification via support vector classifiers, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 105–108.

Senoussaoui, M., Kenny, P., Brummer, N., de Villiers, E., Dumouchel, P., 2011. Mixture of PLDA models in I-vector space for gender independent speaker recognition, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 25–28.

Silovsky, J., Prazak, J., Cerva, P., Zdansky, J., Nouza, J., 2011. Plda-based clustering for speaker diarization of broadcast streams, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 2909–2912.

Stafylakis, T., Kenny, P., Ouellet, P., Perez, J., Kockmann, M., Dumouchel, P., 2013. Text-dependent speaker recognition using PLDA with uncertainty propagation, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 3684–3688.

Steininger, S., Rabold, S., Dioubina, O., Schiel, F., 2002. Development of user-state conventions for the multimodal corpus in SmartKom, in: LREC Workshop on" Multimodal Resources", Las Palmas, Spain.

Stolcke, A., Mandal, A., Shriberg, E., 2012. Speaker Recognition with Region-Constrained MLLR Transforms, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 4397–4400.

Sturim, D., Reynolds, D., Dunn, R., Quatieri, T., 2002. Speaker verification using text-constrained Gaussian mixture models, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, IEEE; 1999. pp. 677–680.

Subramanya, A., Zhang, Z., Surendran, A.C., Nguyen, P., Narasimhan, M., Acero, A., 2007. A generative-discriminative framework using ensemble methods for text-dependent speaker verification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. IV–25.

Toledano, D.T., Hernandez-Lopez, D., Esteve-Elizalde, C., Fierrez, J., Ortega-Garcia, J., Ramos, D., Gonzalez-Rodriguez, J., 2008. BioSec Multimodal Biometric Database in Text-Dependent Speaker Recognition, in: LREC.

Toledo-Ronen, O., Aronowitz, H., Hoory, R., Pelecanos, J., Nahamoo, D., 2011. Towards Goat Detection in Text-Dependent Speaker Verification, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 9–12.

Vogt, R., Sridharan, S., 2008. Explicit modelling of session variability for speaker verification. Computer Speech & Language 22, 17–38.

Vogt, R.J., Lustri, C.J., Sridharan, S., 2008. Factor analysis modelling for speaker verification with short utterances, in: Odyssey Speaker and Language Recognition Workshop, IEEE. pp. 1–4.

Vogt, R.J., Pelecanos, J., Scheffer, N., Kajarekar, S., Sridharan, S., 2009. Within-session variability modelling for factor analysis speaker verification, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 1563–1566.

Wagner, M., Summerfield, C., Dunstone, T., Summerfield, R., Moss, J., 2006. An evaluation of "commercial off-the-shelf" speaker verification systems, in: Odyssey Speaker and Language Recognition Workshop, pp. 1–8.

Wong, Y.W., Chang, S.I., Seng, K.P., Ang, L.M., Chin, S.W., Chew, W.J., Lim, K.H., 2011. A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities. Pattern Recognition Letters 32, 1503 – 1510. doi:10.1016/j.patrec.2011.06.011.

Woo, R.H., Park, A., Hazen, T.J., 2006. The MIT Mobile Device Speaker Verification Corpus: Data Collection and Preliminary Experiments, in: Odyssey Speaker and Language Recognition Workshop.

Woo, S.C., Lim, C.P., Osman, R., 2000. Text-dependent speaker recognition using the fuzzy ARTMAP neural network, in: Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology, TENCON, Kuala Lumpur (Malaysia). doi:10.1109/TENCON.2000.893535.

Wu, D., BaojieLi, Jiang, H., 2008. Speech recognition, technologies and applications - Normalization and transformation techniques for robust speaker recognition. I-Tech, Vienna, Austria.

Xu, J., Zhang, Y., Yan, Z.J., Huo, Q., 2011. An i-vector based approach to acoustic sniffing for irrelevant variability normalization based acoustic model training and speech recognition, in: Annual Conference of the International Speech Communication Association (Interspeech), pp. 1701–1704.

Yegnanarayana, B., Prasanna, S.M., Zachariah, J.M., Gupta, C.S., 2005. Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system. IEEE Transactions on Speech and Audio Processing 13, 575–582.

Yoma, N.B., Pegoraro, T.F., 2002. Robust speaker verification with state duration modeling. Speech Communication 38, 77–88.

You, C., Lee, K.A., Li, H., 2010. GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing 18, 1300–1312.

Young, S.J., 1992. The general use of tying in phoneme-based HMM speech recognisers, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 569–572.

Young, S.J., 2008. Springer Handbook of Speech Processing. Springer-Verlag. chapter HMMs and Related Speech Recognition Technologies. pp. 539–557.

Zheng, T.F., 2005. The voiceprint recognition activities over China, in: Oriental COCOSDA.